UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

An application of extreme value theory in medical sciences

Constantino Pereira Caetano

Mestrado em Bioestatística

Dissertação orientada por:

Professora Doutora Patrícia de Zea Bermudez

2018

Acknowledgements

In no special order, I would like to profoundly thank the following individuals and institutions for helping me realize this lifelong dream.

To my mother, for inspiring me to pursue a scientific career and giving me the possibility to do so.

To my step-father, for the excellent mathematical input into my work.

To my father, with whom I had the most fascinating mathematical discussions that inspired me to pursue this journey.

To my advisor, Professor Patrícia de Zea Bermudez, who presented me with the most interesting challenges throughout this dissertation.

To my friend João Torrado, for giving me a place to stay and being a friend to debate statistics with.

To Dr. Eduardo Gomes da Silva, head of medicine of Serviço de Medicina 3.2 of the Hospital dos Capuchos, Lisbon, who offered an afternoon of his time to answer my questions about blood pressure and its pathologies.

To the Department of Pharmaceutical Care Services of the Portuguese National Association of Pharmacies (ANF), for enabling me the use of the data used in this thesis.

Thank you.

Resumo

Valores altos de pressão arterial são considerados um fator de risco de doenças cardiovasculares, ver [Hajar, 2016]. Estas doenças são a principal causa de morte em Portugal. Com o objectivo de criar um perfil da população Portuguesa em relação aos riscos de doenças cardiovasculares, um estudo foi desenvolvido em 2005, pela Associação Nacional de Farmácias através do seu Departamento de Serviços Farmacêuticos. O interesse principal do presente estudo consiste em modelar valores elevados de pressão arterial sistólica em indivíduos que sofrem de uma categoria particular de hipertensão. Um estudo similar foi desenvolvido para modelar os valores elevados de níveis de colesterol total, ver [de Zea Bermudez and Mendes, 2012]. A presente dissertação tem dois principais interesses: estudar a distribuição geográfica dos valores elevados de pressão arterial sistólica (em indivíduos com valores normais de pressão arterial diastólica) em Portugal, i.e., ajustar modelos de valores extremos para cada distrito de Portugal e ilhas e analisar em particular o grupo de maior risco, i.e., indivíduos idosos. Com esse propósito, a metodologia *Peaks Over Threshold* foi aplicada. Esta metodologia consiste em ajustar um modelo aos excessos (ou excedências) acima de um limiar de pressão arterial sistólica suficientemente elevado. Os modelos obtidos serão capazes de estimar quantis elevados e probabilidades de *cauda* de pressão arterial sistólica.

Na presente dissertação, os indivíduos foram divididos em quatro grupos distintos. Aqueles que apresentavam valores normais de pressão arterial sistólica e pressão arterial diastólica. Os que apresentavam valores superiores aos delineados pelas entidades médicas em um ou ambos os índices, ver tabela 6.1. Dentro deste último grupo consideramos os indivíduos que sofrem de hipertensão arterial sistólica isolada, caracterizada por valores de pressão arterial sistólica superior ou igual a 140 mmHg e valor de pressão arterial diastólica inferior a 90 mmHg. Pretendemos estudar valores elevados de pressão arterial sistólica neste grupo.

Em primeira análise, foi efectuado um estudo descritivo dos indivíduos que frequentaram a campanha e que sofrem de hipertensão sistólica isolada, com o intuito de averiguar o efeito de outras variáveis de interesse nos níveis de pressão arterial sistólica. As variáveis consideradas nesta analise preliminar foram a idade, cuja relação com valores elevados de pressão arterial sistólica é conhecida, ver [Pinto, 2007]; o género, consumo de tabaco, índice de massa corporal e distrito.

A análise de valores extremos utilizando a metodologia *Peaks Over Threshold* consiste em várias etapas. Em primeiro lugar, é necessário obter o valor limiar elevado (threshold) com o objectivo de ajustar uma distribuição generalizada de Pareto aos seus excessos. Esta distribuição tem parâmetro de forma k e parâmetro de escala σ, ver expressão (3.2). Esta primeira etapa é por vezes difícil. A literatura apresenta várias metodologias para tratar esta fase. Existem métodos exploratórios, como o descrito por [Coles, 2001], que utiliza a função de excesso médio para discernir o limiar elevado pretendido. [DuMouchel, 1983] sugere utilizar $\chi_{0.9}$ como valor limiar. Existem também métodos que consistem em ajustar o modelo considerando vários valores limiar e avaliar qual produz o melhor ajustamento, como por exemplo os testes de Cramér-von Mises e Anderson-Darling, ver [Choulakian and Stephens, 2001]. Ainda dentro deste grupo destacamos um método Bayesiano que utiliza medidas de surpresa, ver [Lee et al., 2015]. Todos os métodos referidos acima são utilizados ao longo da dissertação.

Após concluída esta fase procedemos ao ajuste de uma distribuição generalizada de Pareto aos excessos do valor limiar seleccionado. Máxima verosimilhança é a metodologia mais usual para efectuar o

ajustamento visto que os resultantes estimadores dos parâmetros gozam de propriedades relevantes.

Numa primeira etapa, implementamos a metodologia *Peaks Over Threshold* nos indivíduos que sofrem de pressão arterial sistólica isolada em cada distrito de Portugal continental e ilhas. Aqui são exploradas as dificuldades inerentes na análise de valores extremos e também alguns problemas encontrados nos dados, os quais são explorados no capítulo seguinte, onde analisamos os valores de pressão arterial sistólica em indivíduos idosos, (idade superior ou igual a 55) e consideramos um método que trata o problema de testes múltiplos para hipóteses ordenadas. Estas resultam da aplicação dos testes de Cramér-von Mises e Anderson-Darling para diferentes partições da amostra; e consideramos também modelos *jittering* para lidar com o problema de discretização dos dados.

Palavras-chave: Teoria de Valores Extremos, Estatística Bayesiana, Modelos *Threshold*, Escolha de *Threshold*, Testes Múltiplos Para Hipóteses Ordenadas.

Abstract

It has been well stated that high values of blood pressure constitute a risk factor for cardiovascular diseases [Hajar, 2016], with the latter being the number one death cause in Portugal. With the objective of profiling the Portuguese population in what regards cardiovascular diseases' risk factors, a study was developed and carried out in 2005, by the National Pharmacy Association through its Department of Pharmaceutical Care. The main interest of the present study is to model the high values of systolic blood pressure of individuals with a specific hypertension pathology. A similar study was developed for the total cholesterol levels [de Zea Bermudez and Mendes, 2012]. The aims of this dissertation are twofold: to study the geographical distribution of the high systolic blood pressure (but normal diastolic blood pressure) in Portugal, i.e., fitting extreme value models for each Portuguese district and islands and studying the group that is more at risk, i.e., the elderly. With that purpose, the *Peaks Over Threshold* methodology was applied, which consists in finding a sufficiently high systolic blood pressure threshold and fitting a tail model to the excesses. The models will be able to estimate extreme quantiles and tail probabilities of the systolic blood pressure in each group.

Keywords: Extreme Value Theory, Bayesian Statistics, Threshold Models, Threshold Selection, Multiple Testing For Ordered Hypotheses.

Contents

List of figures

List of tables

1 Introduction

Extreme events can be defined as low frequency episodes of some apparently random process. For example, floods transpire when the water level of some water body exceeds an uncommonly high threshold. Classical statistical methodologies are not suited to treat this kind of data, since they aim to make predictions about future behavior of the treated phenomena based on the most common events, i.e., classical statistics uses the centralized data to infer on future behavior by fitting the data to models based on asymptotic central limit like results. Such an approach might be considered too overly simplified to infer on rare events. Hence, the extreme value analysis paradigm arose out of the necessity to treat data that falls on this scope. It offers well suited statistical procedures to describe the tail distribution behavior of the underlying data creation process.

Extreme value theory (EVT) can be applied to an assortment of different scientific and economic branches, ranging from meteorology, hydrology and insurance, amongst others. It was introduced by Leonard Tippett (1902-1985) and Sir Ronald Aylmer Fisher (1890-1962). Tippett worked for the British cotton industry research association where he developed research to make the cotton threads stronger. He showed that the cotton thread is only as strong as its weakest fibers. Along with Fisher, Tippet laid the probabilistic frameworks of what became the extreme value theory as it is known today. Gumbel was also an outstanding contributer to the development of EVT, see [Gumbel, 1935].

Like classical statistics, EVT aims to fit known statistical models to the data. Since the interest is to fit models to the least common data, it is natural to consider models for the tail of the distribution. These were first introduced by Fisher and Tippet. They proved that, under the criteria of independent and identically distributed (i.i.d.) random variables, the sample maxima (or minima) could be modeled by one of three tail distributions of extremes as the sample size increases. This method is widely known in the literature as the *Annual Maxima*.

Other methods include the *Peaks Over Threshold* approach (POT). This methodology aims to fit a tail-like model to the excesses over a high threshold. It was first developed by [Pickands, 1975] and [Balkema and de Haan, 1974]. Research into POT models is an ongoing topic. In this dissertation, we use several recently developed methods that provide further evidence in selecting the most suited threshold model for the data.

One of the most strenuous analysis using POT is the selection of the threshold value, i.e., the value over which the tail-like distribution model is fitted. We address several recent methods pertaining to this analysis, see [Lee et al., 2015], [Bader et al., 2018] and [Coles, 2001].

Models obtained by using EVT techniques are able to extrapolate on the likelihood of future extreme events, even if such events were not observed during the sampling process. For example, if one were to record Lisbon's precipitation levels each day for a long period of time during the rainy season, one might never observe a flood or a dry spell. Though the data could be fitted to an EVT model that could extrapolate the likelihood of observing such events, the interest might also lie in finding the value that is exceeded, in mean, once every rainy season, which the latter also foresees.

Although not common, EVT can be used to treat medical data. Furthermore, some literature exists on this topic, [de Zea Bermudez and Mendes, 2012] fit EVT models to the total cholesterol recorded of Portuguese voluntary pharmacy attendees. In this dissertation we use POT techniques to fit threshold models to systolic blood pressure data. We use a database obtained during a study of risk factors associated with cardiovascular diseases in the Portuguese population, where in certain Portuguese pharmacies, voluntary

attendees had several biometric variables recorded, such as total cholesterol, triglycerides, systolic blood pressure, diastolic blood pressure and body mass index, amongst others.

Hypertensive individuals have a higher risk rate of contracting heart diseases, see [Hajar, 2016]. With the objective of analyzing this health issue in Portugal, a campaign was carried out by the National Pharmacy Association through its Department of Pharmaceutical Care in 2005, with the goal of identifying individuals at risk. This dissertation aims to study individuals who suffer from isolated systolic hypertension (ISH), that is, individuals that have systolic blood pressure higher or equal to 140 mmHg and diastolic blood pressure lesser than 90 mmHg (current guidelines from the Portuguese Hypertension Society).

In the first chapters of this dissertation we outline the main EVT theorems and results. We delve more into the intricacies of the threshold models, where we formulate the POT's asymptotic distribution of excesses (or exceedances) above a sufficiently high threshold. This will be the main methodology applied throughout this dissertation. We then present several methods for threshold selection, namely an exploratory method using the mean residual life function as described by [Coles, 2001] and one method using goodness-of-fit hypotheses testing for the generalized Pareto distribution (GPD), with a rule to account for multiple ordered hypotheses testing, see [Choulakian and Stephens, 2001] and [Bader et al., 2018]. We also include a beginner's guide to Bayesian statistics, since one of the methods proposed for threshold selection requires some knowledge on this domain. This state-of-the-art method uses the predictive p-value obtained by sampling from the predictive posterior distribution to calculate tail-area probabilities by considering a range of threshold candidates. These p-values can be interpreted so that a suited threshold is selected. The R code used in this dissertation for the application of this method was courteously made available by the authors who developed it, see [Lee et al., 2015]. Chapter 5 presents the basic concept of the delta method which will be used in subsequent chapters.

Chapter 6 presents the exploratory analysis of the individuals who volunteered to join the campaign. Several variables were examined alongside systolic blood pressure, such as age, body mass index, gender and district.

Chapter 7 and 8 pertain to the extreme value modeling of systolic blood pressure for the previously mentioned individuals who suffer from ISH. We perform a first approach for each Portuguese district using these methods, highlighting the difficulties involved with the threshold model analysis. On a second stage, we model the systolic pressure of the elderly individuals ($>= 55$) who suffer from ISH, with the objective of applying more rigorous statistical analysis to obtain better suited models. This group is of particular interest, since elderly people have a higher prevalence of ISH [Bavishi et al., 2016].

2 | Models for Extreme Values

2.1 Introduction

In today's modernized and highly technological world, it is imperative to study the risk of extreme events associated with economical and natural disasters, such as the 2008 Wall Street stock market crash and Hurricane Maria that devastated Puerto Rico in 2017. The likelihood of these events is very low, meaning that if they were quantified through standard statistical analysis, they would fall on the tail-end of the distribution.

Let's consider total rainfall as an example. We are interested in assessing the amount of rainfall that is needed to cause a flood and not the usual mild rain. This amount in not usually observed, hence being termed a rare event. Extreme value analysis offers well suited statistical techniques capable of predicting these events. It outlines statistical procedures to predict certain quantities of interest, such as the most extreme value of a given distribution known as *endpoint* (if it is finite), the probability of some extreme value being exceeded, termed *tail probability* and extreme quantile estimation or *tail quantiles*, i.e., values that have a low probability of being exceeded.

In this chapter we explore the asymptotic frameworks of classical extreme theory.

2.2 Classical Extreme Value Theory

Although there is some literature on sequences of dependent variables for extreme value analysis, this setting does not fall under the scope of this thesis. [Coles, 2001] has a chapter dedicated to this topic. Hence we consider the standard sequence of independent variables with common distribution. Let X be a random variable with distribution function F and

$$M_n = \max(X_1, X_2, ..., X_n), \tag{2.1}$$

where $X_1, X_2, ..., X_n$ represents a random sample of X. The aim is to study the distribution of M_n. We can derive this distribution as follows:

$$\begin{aligned} P(M_n \leq z) =& P(X_1 \leq z, X_2 \leq z, ..., X_n \leq z) \\ =& P(X_1 \leq z) \times (X_2 \leq z) \times ... \times P(X_n \leq z) \\ =& \{F(z)\}^n. \end{aligned} \tag{2.2}$$

In practice this is not very helpful since the distribution F is usually not known.

Theorem 2.2.1 (Fisher-Tippet Theorem) *If there exist sequences of constants* $(a_n > 0)$ *and* (b_n) *such that*

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z), \qquad n \to \infty$$

where G is a non-degenerate distribution function, then G belongs to one of the following families:

$$I : G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \quad -\infty < z < \infty;$$

$$II : G(z) = \begin{cases} 0, & z \leq b, \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b; \end{cases}$$

$$III : G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\}, & z < b, \\ 1, & z \geq b; \end{cases}$$

for parameters a>0, $b \in \mathbb{R}$ *and, in the case of families II and III,* $\alpha > 0$*. The parameters a, b and* α *are the scale, location and shape parameters, respectively.* □

Theorem 2.2.1 shows that the distribution of the rescaled sample maximum converges in distribution, as the sample size grows to infinite, to one of the three distribution families, I, II and III also known as Gumbel, Fréchet and Weibull families, respectively. This theorem has a major importance in EVT because it states that no matter the underlying distribution of X, the rescaled M_n has one of the previously mentioned asymptotic extreme distributions.

These distributions can be combined into a single family, termed the generalized extreme value distribution (GEV).

Theorem 2.2.2 (The Generalized Extreme Value Distribution) *If there exist sequences of constants* $(a_n > 0)$ *and* (b_n) *such that*

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z), \qquad n \to \infty$$

where G is a non-degenerate distribution function, then G belongs to the GEV family

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\},$$

with support $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$*, where* $-\infty < \mu < \infty$*,* $\sigma > 0$ *and* $-\infty < \xi < \infty$*.*

See [Coles, 2001] for an outline proof of the GEV theorem. This family will be used in the next chapter as an approximation of the distribution of M_n to deduce the conditional distribution necessary to draft the threshold models.

3 | Threshold Models

3.1 Introduction

In this chapter we will discuss the theoretical fundamentals of the threshold models used in this thesis. Unlike the *Annual Maxima* approach, which uses the distribution of the largest order statistic observed in each block (for instance the maximum monthly temperature recorded in some site), the threshold approach models the data's excesses $(x_i - u)$ above a high value u. This approach is more adequate for non blocked data, such as when only one observation per individual is obtained. Blocked data is usually related to a temporal structure, which may not exist in many applications. In further chapters, we will use this methodology to model the systolic blood pressure (in mmHg) obtained for each Portuguese pharmacy voluntary attendee.

Let $X_1, X_2, ..., X_n, ...$ be a sequence of i.i.d. random variables, each having marginal distribution function F. We intend to find the distribution of the events whose values are higher than a fixed value u, termed threshold. To this intent we consider the following conditional probability

$$P(X > u+x|X > u) = \frac{1-F(u+x)}{1-F(u)}, \quad x > 0 \tag{3.1}$$

where X is an arbitrary term from the $X_1, X_2, ..., X_n, ...$ sequence. If F were known, then the distribution of the excesses could be established. In applications, F is usually not known. Therefore, approximations for a high value of u should be obtained.

3.2 The Generalized Pareto Distribution

Let X be a random variable following a generalized Pareto distribution with shape parameter k, $-\infty < k < \infty$ and scale parameter σ, $\sigma > 0$. X has the following distribution function

$$F(x) = \begin{cases} 1 - \left(1 + \frac{kx}{\sigma}\right)^{-\frac{1}{k}} & k \in \mathbb{R} \setminus \{0\}, \\ 1 - e^{-\frac{x}{\sigma}} & k = 0, \end{cases} \tag{3.2}$$

with support $\left\{x \in \mathbb{R} : x > 0\right\}$ for $k \geq 0$ and support $\left\{x \in \mathbb{R} : 0 < x < -\frac{\sigma}{k}\right\}$ for $k < 0$. The distribution will be denoted by GPD(k,σ).

3.3 The Generalized Pareto Distribution and Threshold Models

Let $X_1, X_2, X_3, ..., X_n, ...$ be a sequence of i.i.d. random variables with distribution function F and

$$M_n = \max(X_1, X_2, X_3, ..., X_n).$$

M_n has GEV distribution function for a large enough n, as stated in theorem 2.2.2, thus

$$F_{M_n}(x) = F^n(x) \approx \exp\left\{ -\left[1 - k\left(\frac{x-\mu}{\sigma}\right)\right]^{\frac{1}{k}} \right\},$$

for certain parameters μ, $\sigma > 0$ and k. If we apply the natural logarithm to both terms, we obtain

$$n \ln(F(x)) \approx -\left[1 - k\left(\frac{x-\mu}{\sigma}\right)\right]^{\frac{1}{k}}.$$

The linear Taylor expansion of $\ln(x)$ is defined as

$$\ln(x) \approx \ln(a) + \frac{1}{a}(x-a)$$

for $a > 0$. Let $a = 1$ then

$$\ln(F(x)) \approx (F(x) - 1).$$

Hence

$$1 - F(x) \approx \frac{1}{n}\left[1 - k\left(\frac{x-\mu}{\sigma}\right)\right]^{\frac{1}{k}}.$$

Computing $1 - F(u+x)$ and $1 - F(u)$ given by the previous approximation onto (3.1) we obtain

$$\begin{aligned} P(X > u+x | X > u) &\approx \frac{\frac{1}{n}\left[1 - k\left(\frac{u+x-\mu}{\sigma}\right)\right]^{\frac{1}{k}}}{\frac{1}{n}\left[1 - k\left(\frac{u-\mu}{\sigma}\right)\right]^{\frac{1}{k}}} \\ &= \left[\frac{1 - k\left(\frac{u+x-\mu}{\sigma}\right)}{1 - k\left(\frac{u-\mu}{\sigma}\right)}\right]^{\frac{1}{k}} \\ &= \left[\frac{1 - k\left(\frac{u-\mu}{\sigma}\right) - \frac{kx}{\sigma}}{1 - k\left(\frac{u-\mu}{\sigma}\right)}\right]^{\frac{1}{k}} \\ &= \left[1 - \frac{kx}{\sigma - k(u-\mu)}\right]^{\frac{1}{k}}. \end{aligned}$$

Letting $\tilde{\sigma} = \sigma - k(u-\mu)$ we obtain

$$\begin{aligned} P(X > u+x | X > u) &\approx \left(1 - \frac{kx}{\tilde{\sigma}}\right)^{\frac{1}{k}}, \\ P(X - u < x | X > u) &\approx 1 - \left(1 - \frac{kx}{\tilde{\sigma}}\right)^{\frac{1}{k}}, \end{aligned} \tag{3.3}$$

thus arriving at the GPD. See [Coles, 2001] for theoretical examples of threshold exceedance models.

3.4 Methods for Threshold Selection

The previous result lays a framework to model excesses over a high threshold u. Given $x_1, x_2, ..., x_n$, a sample derived from i.i.d. random variables, let $x_{1_u}, x_{2_u}, ..., x_{m_u}$ be the data consisting of values larger than u. We intend to fit a GPD to the excesses $x_{i_u} - u$, for $i = 1, ..., m$, which are also considered realizations of i.i.d. random variables.

Choosing an adequate threshold is a difficult task. Selecting too low a threshold may violate the asymptotic basis of the model, which leads to bias. If the threshold is too high it will lead to high variance, which will result in poor estimations of the parameters and extreme quantiles. An adequate threshold should be the lowest value that still provides an acceptable model approximation. Several methods are available for choosing the threshold.

Exploratory methods can be applied prior to the model estimation. Goodness-of-fit tests, which measure the fit of the data to the GPD, can also help in the selection of an adequate threshold. See [Scarrott and MacDonald, 2012] for an outline of several rules and methods for threshold selection. The following subsections will address some of these methods.

3.4.1 Mean Residual Life Function

We intend to fit a generalized Pareto distribution to the excesses $x_{i_u} - u$ above some high threshold u, hence the resulting GPD will yield a null location parameter. Let Y be a random variable that has a generalized Pareto distribution with scale parameter $\sigma > 0$, shape parameter $k \in \mathbb{R}$ and location parameter $\mu = 0$, then

$$E(Y) = \frac{\sigma}{1-k}, \tag{3.4}$$

for $k < 1$. This mean value is infinite for $k \geq 1$. Applying this mean value to the threshold models framework, we obtain

$$e(u_0) = E(X - u_0 | X > u_0) = \frac{\sigma_{u_0}}{1-k}, \tag{3.5}$$

for some threshold u_0 selected from the sequence $X_1, ..., X_n, ...$ of which X is a random term. As stated in the previous section, if a GPD is suited for some threshold u_0, then it is also valid for some threshold $u > u_0$. Hence

$$E(X - u | X > u) = \tfrac{\sigma_u}{1-k}$$

From (3.3) we can deduce that $\sigma_u = \sigma_{u_0} - k\,u$, thus

$$E(X - u | X > u) = \frac{\sigma_{u_0} - k\,u}{1-k} \tag{3.6}$$

Equation (3.6) suggests that $E(X - u | X > u)$ is a linear function of u with intercept $\frac{\sigma_0}{1-k}$ and slope $-\frac{k}{1-k}$. Let $x_1, x_2, ..., x_n$ be a given sample, where x_{i_u}, $i = 1, ..., m$ are the values of the sample that are greater than u. The mean excess function is given by

$$\hat{e}(u) = \frac{1}{n_u} \sum_{i=1}^{n_u} \left(x_{i_u} - u \right), \quad u < \max(x_1, x_2, ..., x_n) \tag{3.7}$$

This function provides an empirical estimate of $E(X - u | X > u)$, hence plotting $(u, \hat{e}(u))$ should provide a linear function for some high threshold u.

Let's consider the following example: the data was simulated from the mixture distribution $X_{sim} \sim 0.3\,U(0, 20) + 0.7\text{GPD}(u = 20, \sigma = 18, k = -0.07)$, $n = 1000$. We plot the mean residual life function for $u = 1, ..., 100$, obtaining the following representation:

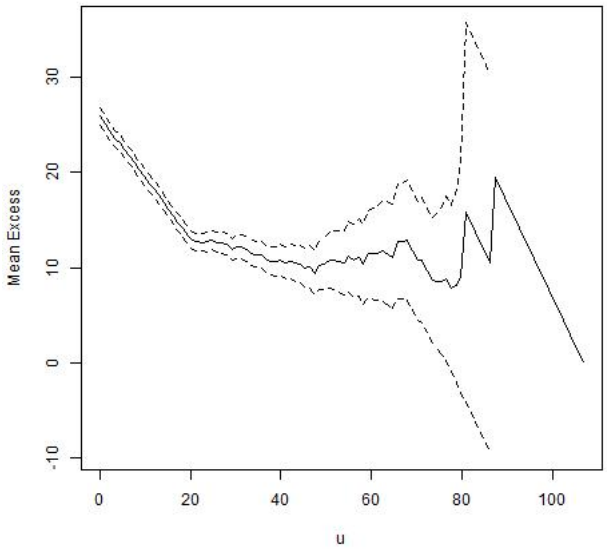

Fig. 3.1: Mean residual life plot for the dataset simulated from a mixture distribution

The plot shows that there is a clear change in the function's behavior for $u = 20$, alluding that 20 is the adequate threshold. This method can be computed without the need to fit a GPD to the data, which is an upside. Being a graphical diagnostic technique, it has the downside of being hard to interpret in some circumstances.

3.4.2 Goodness-of-fit Tests for the Generalized Pareto Distribution

In this section we present a summarized description of the procedure using the Cramér-von Mises (W^2) and Anderson-Darling (A^2) goodness-of-fit tests for the generalized Pareto distribution to ascertain or select an adequate threshold, see [Choulakian and Stephens, 2001] and [Bader et al., 2018], while accounting for multiple testing error control as described by [Bader et al., 2018]. The null hypothesis is H_0 : the random sample $x_1, x_2, ..., x_n$ comes from a generalized Pareto distribution. The Anderson-Darling statistic is a modification of the Cramér-von Mises statistic that gives more weight to the observations in the tail of the distribution. Their test statistics are as follows:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\big[\ln(z_i) + \ln(1 - z_{n+1-i})\big], \tag{3.8}$$

$$W^2 = \sum_{i=1}^{n}\big[z_i - \frac{(2i-1)}{(2n)}\big]^2 + \frac{1}{12n}, \tag{3.9}$$

where $z_i = F(x_i)$ and F is the distribution function of the GPD.

Several cases are outlined in [Choulakian and Stephens, 2001] regarding the amount of information about the parameters of distribution. In the case study presented in a further chapter both σ and k are unknown, and so an estimation of these is desired. The outline of the method is as follows:

1. Calculate estimates of k and σ via maximum likelihood and compute $z_i = F(x_i)$ for $i = 1, ..., n$.
2. Calculate the test statistics A^2 and W^2 given in (3.8) and (3.9)

See [Choulakian and Stephens, 2001] for the tables containing asymptotic percentiles of W^2 and A^2 for the case where both σ and k are unknown.

The aforementioned method can be used to ascertain the quality of the fit for a range of threshold candidates. The outlining of this procedure is as follows:

1. Select an array of sorted threshold candidates $u_1 < u_2 < ... < u_m$.
2. Set a proper significance level α.
3. For each u_i, $i = 1, ..., m$ select the excesses y_j, $j = 1, ..., l_i$.
4. Compute W^2 and A^2 for each u_i, $i = 1, ..., m$.
5. Select the lowest value of u_i for which the null hypothesis is not rejected for the significance level chosen in 2.

It is not desired to keep the null hypothesis at a low threshold. This problem can happen by chance, hence procedures to avoid this situation are necessary. These tests are ordered, meaning that if H_0^i is rejected for some i, all other H_0^k, $1 \leq k < i$ must also have been rejected.

We now outline the FowardStop rule addressed by [Bader et al., 2018] and first proposed in the literature by [G'Sell et al., 2015] to handle ordered multiple hypotheses testing. Let's consider a sequence of null hypothesis H_0^1,...,H_0^m, where H_0^i: the excesses over u_i come from a generalized Pareto distribution with parameter vector (k,σ), where k and σ should be estimated. Let $p_1, p_2, ..., p_m$, be the resulting p-values for each test. The idea is to create a function dependent on the previous p-values, which will return a value that can be compared to a pre-specified significance level α. The function considered is the mean of a ln transformation of these p-values, giving more weight to higher p-values and, conversely, giving less weight to smaller p-values. Finally, we select the highest cutoff $\hat{i} \in \{1, ..., m\}$ such that the returned value is still lower than α. The rule can be formulated as follows:

$$\hat{i} = \max \left\{ i \in \{1, ..., m\} : -\frac{1}{i} \sum_{j=1}^{i} \ln(1 - p_j) \leq \alpha \right\}. \tag{3.10}$$

By choosing $\hat{i}$ as the cutoff, we reject H_0^1,...,$H_0^{\hat{i}}$, hence selecting the threshold associated with the $H_0^{\hat{i}+1}$ hypothesis. Several other rules have been outlined by [Bader et al., 2018], such as the StrongStop.

3.4.3 Bayesian Method for Threshold Selection

In the present section, we outline the Bayesian procedure formulated by [Lee et al., 2015], using measures of surprise to quantify the level of incompatibility of the observed data to a generalized Pareto distribution. The fundamentals of Bayesian statistics are briefly reviewed in Chapter 5. [Lee et al., 2015] also outlines methods for threshold selection for the bivariate case. The latter falls outside the scope of this thesis and will therefore not be subsequently referenced. In classical statistics, the p-value is considered a measure of surprise. It measures the likelihood of observing a test statistic as extreme or more extreme than the one observed, under the null hypothesis. The goal of the method proposed by [Lee et al., 2015] is to compute the predictive p-value for a dataset in order to quantify the degree to which a GPD can be adequately fitted. The procedure is as follows: consider a sample $\mathbf{x} = (x_1, x_2, ..., x_n)$, a realization of i.i.d. random variables,

1. Select an array of sorted threshold candidates $u_1 < u_2 < ... < u_m$ in $\{x_1, x_2, ..., x_n\}$. A rule of thumb here is choosing equally distanced values, starting from the lowest (which can be the minimum observed value), such that between each candidate there is a substantial amount of data.
2. For each u_i, $i = 1, ..., m$, build the posterior distribution $\pi(\theta | x_{u_i})$, where $x_{u_i} = \{x_j \in x : x_j > u_i\}$ and $\theta = (k, \sigma)$. A Jeffreys' prior is considered as the distribution for θ [Lee et al., 2015]. If an analytical solution is not possible to derive, the posterior distribution can be obtained by means of numerical approximation or by simulation. Monte Carlo Markov Chains algorithms (MCMC) are simulation algorithms that can be used to sample from the posterior distribution. One such algorithm is the Metropolis-Hastings. This algorithm is outlined in [Turkman et al., 2018].

3. Obtain draws from each $\pi(\theta|x_{u_i})$, $i = 1,2,..,m$. This can be achieved by simulating a sample from each posterior distribution.

4. Using the draws obtained from the previous step, we can input the retrieved values onto the GPD model and generate samples $x_{u_{i_{sim}}}$ from the predictive posterior distributions $m(y_{u_i}|x_{u_i})$.

5. Compute the predictive p-value p_{m_0} given by $P_m(T(X) \geq T(x_{u_{i_{sim}}}))$ using an appropriate test statistic T. Repeat for each candidate threshold. The likelihood was the selected test statistic.

6. For each $i = 1,2,...,m$, plot $(u_i, p_m(i))$.

[Meng, 1994] showed that the expected value of the posterior predictive p-value is 0.5 under the null hypothesis. Hence, predictive p-values closer to 0 or 1 suggest high incompatibility of the data with the null hypothesis H_0 : the sample $x_1, x_2, ..., x_n$ has a generalized Pareto distribution, while p-values closer to 0.5 show less incompatibility with the hypothesis. Note that no alternative hypothesis is specified, see [Lee et al., 2015]. The aim is to find the lowest threshold value that produces a predictive p-value close to 0.5. The choice of the test statistic in step 5 is addressed in [Lee et al., 2015], where the authors present several alternatives and produce examples of their measuring capabilities. As previously mentioned, the R code used in this dissertation for the application of the method was courteously made available by the authors, see [Lee et al., 2015].

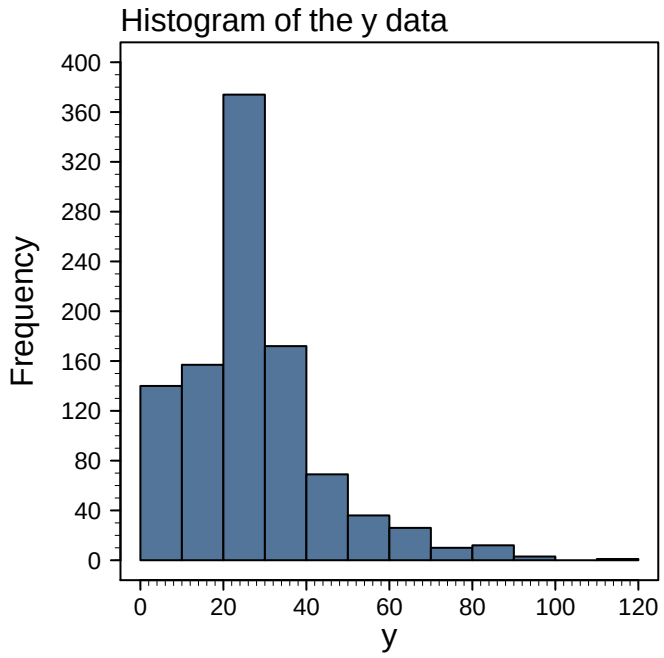

Fig. 3.2: Histogram of the mixture model

Let's illustrate this method. By examining the following mixture distribution $Y \sim 0.3U(0,20) + 0.7\exp(\sigma = 15, \mu = 20)$, a sample of size 1000 is generated from this mixture. The histogram of the data is presented in Figure 3.2. Figure 3.3 shows the output of the Bayesian method when applied to the generated sample at an array of thresholds $(0,5,10,...,40)$. Just as carried out when considering the MEF, we set the threshold at 20. The output clearly shows that $u = 20$ is the lowest threshold such that the resulting predictive p-value is closer to 0.5. This method will be used in subsequent chapters to ascertain an adequate threshold for the systolic blood pressure in Portuguese voluntary pharmacy attendees.

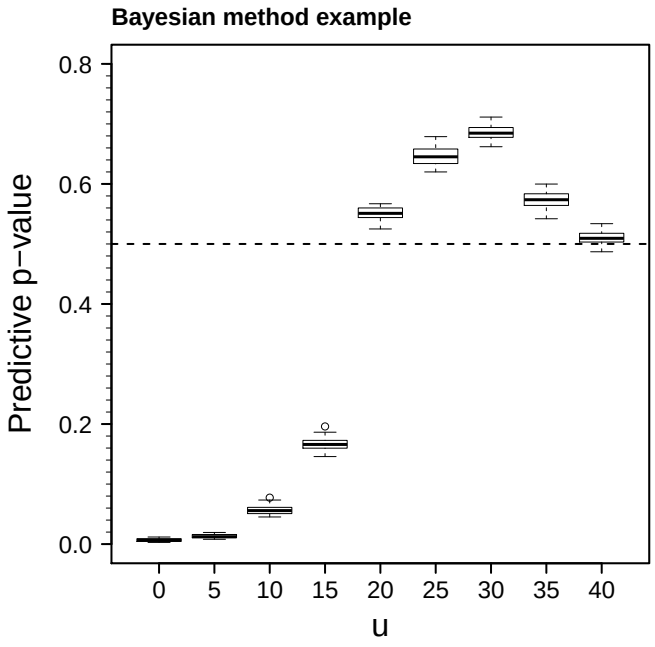

Fig. 3.3: Plotted predictive p-values from a mixture distribution for an array of ordered thresholds

3.5 The Peaks Over Threshold Methodology

Depending on the type of the observations, there are specific methods outlined in the literature to handle extreme value data, see e.g. [Pickands, 1975] and [Coles, 2001]. In this section we focus on the *Peaks Over Threshold* methodology, which uses the exceedances x_{i_u}, $i = 1,...,n$ that fall beyond a high value u to fit a generalized Pareto distribution to the excesses $x_{i_u} - u$ for some given sample $x_1, x_2, ..., x_N$, N is the sample size, while n is the number of exceedances. Most frequently, the parameters of the GPD are estimated by maximum likelihood (ML), namely due to the favorable properties of this method, [Casella and Berger, 2002].

3.5.1 Maximum Likelihood Estimation

Suppose that the values $z_1, z_2, ..., z_n$ are the excesses for a given threshold u. We can obtain the likelihood function for the generalized Pareto distribution from (3.2). Let Z have a GPD distribution with shape parameter k and scale parameter σ then for $k \in \mathbb{R} \setminus \{0\}$

$$\frac{dF(z)}{dz} = \frac{1}{\sigma}\left(1 - \frac{kz}{\sigma}\right)^{\frac{1}{k}-1}, \tag{3.11}$$

thus

$$L(\sigma, k; z_1, z_2, ..., z_n) = \frac{1}{\sigma^n}\prod_{i=1}^{n}\left(1 - \frac{kz_i}{\sigma}\right)^{\frac{1}{k}-1}. \tag{3.12}$$

Applying the natural logarithm to both sides, we obtain

$$\begin{aligned} l(\sigma, k; z_1, z_2, ..., z_n) &= \ln\left(\frac{1}{\sigma^n}\right) + \sum_{i=1}^{n}\ln\left[\left(1 - \frac{kz_i}{\sigma}\right)^{\frac{1}{k}-1}\right] \\ &= -n\ln(\sigma) + \left(\frac{1}{k} - 1\right)\sum_{i=1}^{n}\ln\left(1 - \frac{kz_i}{\sigma}\right), \end{aligned} \tag{3.13}$$

so that $(1 - k z_i/\sigma) > 0$ for $i = 1, ..., n$. In the case where $k = 0$ we can derive the likelihood function as follows

$$\frac{dF(z)}{dz} = \frac{1}{\sigma} e^{-\frac{z}{\sigma}},$$

hence

$$L(\sigma; z_1, z_2, ..., z_n) = \frac{1}{\sigma^n} \prod_{i=1}^{n} e^{-\frac{z_i}{\sigma}},$$

applying the natural logarithm to both sides, we obtain

$$l(\sigma; z_1, z_2, ..., z_n) = -n \ln(\sigma) - \frac{1}{\sigma} \sum_{i=1}^{n} z_i. \tag{3.14}$$

The usual analytical procedure for the maximization of the log-likelihood function does not provide separate expressions for the estimates of k and σ in the case where the former differs from 0. Hence, a numerical technique is required, see [Grimshaw, 1993].

There is an abundance of literature on estimating the parameters of the GPD by other methods besides ML, such as the following: [Castillo and Hadi, 1997], [de Zea Bermudez and Kotz, 2010a] and [de Zea Bermudez and Kotz, 2010b]. Maximization of the previous functions with respect to the parameters (σ, k) produces maximum likelihood estimators $(\hat{\sigma}, \hat{k})$, which have been shown to be asymptotically normal and asymptotically efficient under certain conditions of regularity. These conditions are needed, since when $k > 0$, the support of the generalized Pareto distribution is dependent on its parameters.

For $k = 0$, the GPD is reduced to the exponential distribution. Simple derivation methods are sufficient to obtain maximum likelihood estimators in this case. Let $Z \sim Exp(\sigma)$ and $z_1, z_2, ..., z_n$ be a sample derived from Z. Then, by maximizing (3.14), we obtain

$$\begin{aligned}
&\frac{dl(\sigma; z_1, z_2, ..., z_n)}{d\sigma} = 0, \\
&-\frac{n}{\sigma} + \frac{1}{\sigma^2} \sum_{i=1}^{n} z_i = 0 \\
&\Leftrightarrow \frac{1}{\sigma} \left(-n + \frac{1}{\sigma} \sum_{i=1}^{n} z_i \right) = 0 \\
&\Leftrightarrow \frac{1}{\sigma} = \frac{n}{\sum_{i=1}^{n} z_i} \\
&\hat{\sigma} = \frac{\sum_{i=1}^{n} z_i}{n},
\end{aligned} \tag{3.15}$$

then, the estimator of sigma is $\hat{\sigma} = \overline{Z}$.

Hence, the mean is a natural estimator of σ. It is easy to derive confidence intervals for σ by the asymptotic properties of ML estimators.

For $k \neq 0$, [Castillo and Hadi, 1997] outlines an algorithm to compute statistically efficient estimators for k and σ that rely only on two distinctive order statistics in a random sample. The authors also lay methodology out to derive confidence intervals for k and σ using the delta method.

3.5.2 Estimation of Extreme Quantiles

The topic of extreme quantile estimation has been thoroughly studied in the literature, such as [Coles, 2001], [Grimshaw, 1993] and [Hosking and Wallis, 1987]. In this chapter, we summarize the method described in [Coles, 2001] for extreme quantile estimation. Consider that the generalized Pareto distribution (for some unknown parameters $k \neq 0$ and σ) is suited to model the excesses $x - u$ over a threshold u. From (3.3), we obtain

$$P(X > x|X > u) = \left[1 - k\left(\frac{x-u}{\sigma}\right)\right]^{\frac{1}{k}}, \quad x > u, \tag{3.16}$$

also

$$\begin{aligned} P(X > x|X > u) &= \frac{P(X > x, X > u)}{P(X > u)} \\ \Leftrightarrow P(X > x|X > u) &= \frac{P(X > x)}{P(X > u)} \\ \Leftrightarrow P(X > x) &= P(X > x|X > u) \times P(X > u), \end{aligned}$$

thus

$$P(X > x) = \left[1 - k\left(\frac{x-u}{\sigma}\right)\right]^{\frac{1}{k}} \times P(X > u).$$

Let $\tau_u = P(X > u)$. The extreme quantile, which is the value x_p such that $P(X > x_p) = p$ for a very small p, is obtained as follows:

$$\begin{aligned} &\left[1 - k\left(\frac{x_p - u}{\sigma}\right)\right]^{\frac{1}{k}} \times \tau_u = p, \\ \Leftrightarrow\ &1 - k\left(\frac{x_p - u}{\sigma}\right) = \left(\frac{p}{\tau_u}\right)^k, \\ \Leftrightarrow\ &x_p = \frac{\sigma}{k}\left[1 - \left(\frac{p}{\tau_u}\right)^k\right] + u, \end{aligned} \tag{3.17}$$

where it is necessary that p is a probability close to 0 to ensure that $x_p > u$. For $k = 0$, we obtain x_p applying a similar procedure:

$$1 - (1 - e^{-\frac{x-u}{\sigma}}) = \frac{P(X > x)}{P(X > u)},$$

$$P(X > x) = e^{-\frac{x-u}{\sigma}} \times \tau_u, \quad \tau_u = P(X > u),$$

then the extreme quantile with probability $1 - p$ is obtained as follows:

$$\tau_u e^{-\frac{x_p - u}{\sigma}} = p,$$

hence

$$x_p = \sigma \ln\left(\frac{\tau_u}{p}\right) + u. \tag{3.18}$$

Note that in both cases we are estimating the $(1-p)$th quantile, for a very small probability p.

Extreme quantile estimation is obtained by imputation of the maximum likelihood estimates of the GPD parameters. An estimate of τ_u is also necessary, as τ_u is the probability that the random variable X exceeds the threshold u. An obvious estimator of τ_u is

$$\hat{\tau}_u = \frac{M}{n},$$

where M is the amount of order statistics that exceed u and n is the sample size. The number of order statistics exceeding u has a binomial distribution, $\text{Bin}(n, \tau_u)$, where $\hat{\tau}_u$ is also the maximum likelihood estimate of τ_u. Therefore, the uncertainty of τ_u should also be taken into account.

In the case where $k \neq 0$, in order to calculate confidence intervals for the quantiles, it is necessary to calculate the variance-covariance matrix of $(\hat{\sigma}, \hat{k}, \hat{\tau}_u)$.

M has distribution $\text{Bin}(n, \tau_u)$ and let $\hat{\tau_u} = \frac{M}{n}$, where M is the number of upper order statistics that exceed u. We can derive the variance of $\hat{\tau_u}$ as follows:

$$var\left(\frac{M}{n}\right) = \frac{1}{n^2} var(M) = \frac{1}{n^2}[n\tau_u(1-\tau_u)] = \frac{1}{n}[\tau_u(1-\tau_u)]$$

The variance-covariance matrix of $(\hat{\sigma}, \hat{k}, \hat{\tau_u})$ is given by

$$V = \begin{bmatrix} \frac{1}{n}[\tau_u(1-\tau_u)] & 0 & 0 \\ 0 & v_{11} & v_{12} \\ 0 & v_{21} & v_{22} \end{bmatrix},$$

where v_{ij} are the terms of the variance-covariance matrix of $(\hat{\sigma}, \hat{k})$. These are difficult to obtain. In [Castillo and Hadi, 1997], the authors outline an algorithm to obtain efficient estimators for σ and k, and also calculate their variance-covariance matrix.

Finally, the standard error for $\hat{x_p}$ can be obtained via the delta method where the gradient vector of x_p is as follows:

$$\nabla x_p = \begin{bmatrix} \frac{\partial x_p}{\partial \tau_u} \\ \frac{\partial x_p}{\partial \sigma} \\ \frac{\partial x_p}{\partial k} \end{bmatrix} = \begin{bmatrix} \sigma\, p^k\, \tau_u^{-(k+1)} \\ k^{-1}(1-(p/\tau_u)^k) \\ -\sigma\, k^{-2}(1-(p/\tau_u)^k) - \sigma\, k^{-1}(p/\tau_u)^k \ln(p/\tau_u) \end{bmatrix}$$

Thus, by the delta method, we obtain

$$var(\hat{x_p}) \approx \nabla x_p^T\, V\, \nabla x_p.$$

Chapters 4 and 5 present a summary of the delta method and an introduction to Bayesian statistics, respectively. These chapters were created so the reader could follow the implementation of both methodologies throughout this dissertation. Those who are acquainted with these domains can forgo the reading of the chapters mentioned above.

4 | Basics of Bayesian Statistics

In this chapter we present the fundamentals of Bayesian statistics. The objective is to produce a brief account of Bayesian methodologies so the reader can properly interpret the method presented in chapter 3.

The classical statistics approach consists in obtaining a sample $x_1, x_2, ..., x_n$ produced by a random variable X. X has unknown distribution family Q_θ and unknown fixed parameter vector θ. The aim is to find a suitable model for X and, thus, estimate θ generally via maximum likelihood. Inference on the parameters is performed by calculating point estimation, confidence intervals and hypothesis testing for θ.

In this setting, the data is considered to be obtained through a random process, where $x_1, x_2, ..., x_n$ is a realization of the series $X_1, X_2, ..., X_n$, and X_i $(i = 1, .., n)$ and X have common distribution. Hypothesis testing is carried out by considering a null hypothesis H_0 and a test statistic T following some sampling distribution. We then calculate the probability that under H_0 the test statistic produces a value as extreme or more extreme than the one calculated from the observed data. This probability is called the p-value. If the p-value is very low (close to 0), we reject the null hypothesis at some significance level. This means that our choice is not only based on the observed data, but also on data that was not observed. In this scope, it is incorrect to consider the p-value a true probabilistic measure of the likelihood of H_0, because the distribution parameters on which H_0 depends are fixed quantities. We will show that in the Bayesian setting it is possible to obtain such measures.

The Bayesian layout consists in choosing a model, $f(x|\theta)$, for the observed data and a model for θ, $\pi(\theta)$. The latter is termed prior probability distribution, meaning the distribution of θ before the data is observed. We then construct the posterior distribution of θ, $\pi(\theta|x)$, using Bayes' theorem. We can then derive credible regions for θ. These can be interpreted as true probabilistic measures of the likelihood of the θ being contained in such regions. Using the posterior, distribution we can derive the distribution of future observations, termed the posterior predictive distribution. Hypothesis testing is performed via the Bayes' factor where we compute the chances of two distinct hypotheses.

The Bayesian framework offers the advantage of combining known information about the parameter with observed data. This known information about the parameter is incorporated via the prior distribution. High variance prior distributions can be considered when no prior knowledge about the parameter is available, i.e., vague or non-informative prior distributions. A major advantage of Bayesian analysis is that the posterior distribution can be updated as new data becomes accessible, which makes this a great methodology for processing data which becomes available sequentially. This way, a model can be updated as new data is collected. As the $(i+1)$th dataset becomes available, the ith posterior becomes the $(i+1)$th prior.

The Bayesian sphere does not rely on asymptotic theory. There is no distinction in the procedure for small or large samples. Informative priors become more important when the available data is scarce.

The freedom in choosing the prior distribution for the model's parameters might be a downside if not addressed correctly. The posterior distributions are influenced by the selected prior. Hence, if no information is available, one should choose a prior that yields the least amount of information.

This methodology comes with huge computational burden due to the usual numerical costs of calculating the posterior distribution. As an example, the Bayesian method for threshold selection using measures of surprise required one to two hours of computational effort to process information about

8174 individuals, using an i7 intel core processor.

4.1 Bayes' Theorem

In this section we present the Bayes' theorem and how it is derived in the discrete and continuous case. Let A and B be two events (events are sets with measure P). The conditional probability of A given B is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0.$$

Using the law of total probability, we can write $P(B) = P(A \cap B) + P(\bar{A} \cap B)$. Thus

$$P(A|B) = \frac{P(B|A)P(A)}{P(A \cap B) + P(\bar{A} \cap B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}.$$

4.1.1 The Discrete Case

Let X be a random variable where $\mathbf{x}$ is the observed data (of size n) and $f(x|\theta)$ is an appropriate model for it. We consider $\pi(\theta)$ the prior distribution of θ which is the parameter vector with a discrete support. θ can only take values in $\{\theta_1, \theta_1, ..., \theta_m\}$. Using Bayes' theorem, we can write

$$\pi(\theta_i|x) = \frac{f(x|\theta_i)\pi(\theta_i)}{\sum_{j=1}^{m} f(x|\theta_j)\pi(\theta_j)}, \; i = 1, ..., m,$$

where $\pi(\theta_i|x)$ is the conditional distribution of θ_i given x, also known as the posterior distribution of θ_i.

4.1.2 The Continuous Case

Without loss of generality, we consider the case where θ is an unidimensional parameter and has a continuous distribution $\pi(\theta)$ with support

$$\Theta = \{\theta : \pi(\theta) > 0\},$$

and let X be a random variable with density function $f(x|\theta)$ that belongs to the family of distributions $\Upsilon = \{f(x|\theta) : \theta \in \Theta\}$. Now, let $(x_1, x_2, ..., x_n)$ be a realization of the series of i.i.d. random variables $(X_1, X_2, ..., X_n)$ that have common distribution with X. Then, using Bayes' theorem, we can write

$$\pi(\theta|x_1, x_2, ..., x_n) = \frac{f(x_1, x_2, ..., x_n|\theta)\pi(\theta)}{\int_\Theta f(x_1, x_2, ..., x_n|\theta)\pi(\theta)\, d\theta}, \; \theta \in \Theta$$

$$\Leftrightarrow \pi(\theta|x_1, x_2, ..., x_n) = \frac{\prod_{i=1}^{n} f(x_i|\theta)\pi(\theta)}{\int_\Theta \prod_{i=1}^{n} f(x_i|\theta)\pi(\theta)\, d\theta}, \; \theta \in \Theta.$$

Let $L(\theta|x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta)$ be the likelihood function of the observed data. We can write the previous relation as

$$\pi(\theta|x_1, x_2, ..., x_n) \propto L(\theta|x_1, x_2, ..., x_n) \times \pi(\theta),$$

since

$$E_\theta(L(\theta|x_1, x_2, ..., x_n)) = \int_\Theta \prod_{i=1}^{n} f(x_i|\theta)\pi(\theta)\, d\theta$$

is the normalizing constant.

4.2 Predictive Posterior Distribution

Irrespectively of the approach which is used (classical or bayesian), one of the purposes of data modeling is the prediction of future observations. In a Bayesian framework, the prediction of future observations is performed by means of the predictive distribution. The distribution of a future observation y is given by

$$m(y|x) = \int_{\Theta} f(y|\theta)\pi(\theta|x)\, d\theta,$$

where $\mathbf{x} = (x_1, x_2, ..., x_n)$ is the observed data. $\pi(\theta|x)$ represents the posterior distribution of θ and $f(y|\theta)$ represents the model for the future data y. In the formula written above, y and x are considered to be conditionally independent given θ. This distribution is not θ-dependent, since the parameter is integrated out.

4.3 Bayes' Factor

Similar to classical hypothesis testing, in the Bayesian sphere, we consider a null hypothesis H_0: $\theta \in \Theta$. The key difference is that θ has known prior and posterior distributions, unlike classical statistics, where the parameters are fixed. Hence, it is possible to answer the question: what is the probability of H_0? This question can be further divided. What is the probability of H_0 *a priori*, i.e., before seeing the data? What is the probability of H_0 *a posteriori*? What is the probability of H_0 against some other hypothesis H_1? In this section we present a methodology that will produce answers to these questions.

Let $\theta \in \Theta$ be the parameter we intend to infer on. Let $\pi(\theta)$ and $\pi(\theta|x)$ be the prior and the posterior distribution of θ, respectively. We consider the null hypothesis H_0: $\theta \in \Theta_0$ *vs* H_1: $\theta \in \overline{\Theta_0}$. The prior odds of H_0 against H_1 is defined as

$$O(H_0, H_1) = \frac{P(H_0)}{P(H_1)}.$$

The posterior odds of H_0 against H_1 can be obtained similarly

$$O(H_0, H_1|x) = \frac{P(H_0|x)}{P(H_1|x)}.$$

Considering that θ has a continuous distribution, then

$$O(H_0, H_1) = \frac{\int_{\Theta_0} \pi(\theta) d\theta}{\int_{\overline{\Theta_0}} \pi(\theta) d\theta},$$

and

$$O(H_0, H_1|x) = \frac{\int_{\Theta_0} \pi(\theta|x) d\theta}{\int_{\overline{\Theta_0}} \pi(\theta|x) d\theta}.$$

Low values (close to 0) of $O(H_0, H_1)$ suggest that H_1 is more likely than H_0, while the converse suggests that H_0 is more likely than H_1. The interpretation of the posterior odds is the same. Bayes' factor of H_0 against H_1 is given by

$$B = \frac{O(H_0, H_1|x)}{O(H_0, H_1)}.$$

It has similar interpretation to the prior and posterior odds. High values of B point towards H_0 while low values (close to 0) favor H_1. B measures to what extent the strength of the evidence changed our belief in H_0 against H_1.
Table 4.1 presents the interpretations of the values obtained from the Bayes' factor, see [Kass and Raftery, 1995].

Table 4.1: Bayes' factor output interpretation by [Kass and Raftery, 1995]

B	**Strength of the evidence in favor of H_0**
1 to 3	not worth more than a bare mention
3 to 20	positive
20 to 150	strong
>150	very strong

4.4 Predictive p-values

Traditional p-values can be interpreted as the probability of obtaining test statistic values as large as or greater than the one observed under the null hypothesis, thus calculating some probability area of the distribution's tail. Something similar can be obtained in the Bayesian framework. The predictive p-value can be computed using the following condition

$$p_m = P(T(y) \geq T(\mathbf{x})|\mathbf{x}, H_0), \quad H_0 : \theta \in \Theta_0,$$

where y are future observations, i.e., data generated from the predictive posterior distribution, $\mathbf{x}$ is the observed data and T a given test statistic. If T is free of nuisance parameters, then p_m can be calculated. [Meng, 1994] addresses the topic of calculating the predictive p-value when T includes nuisance parameters.

We now outline a bootstrap procedure to compute an estimate of the predictive p-value:

1. Calculate the predictive posterior distribution $m(y|\mathbf{x})$.
2. Generate $y_i = y_{i_1}, y_{i_2}, ..., y_{i_m}$, $i = 1, ..., n$, from $m(y|\mathbf{x})$.
3. Choose a test statistic T.
4. Compute T for each y_i and for the observed data $\mathbf{x}$.
5. The predictive p-value is given as: $p_m = \frac{k}{n}$, where $k = \#\{T(y_i) > T(\mathbf{x}), i = 1, ..., n\}$.

[Meng, 1994] showed that if H_0 holds, then the predictive p-value has expected value 0.5.

5 | The Delta Method

Given a random variable X, it is sometimes necessary to obtain the distribution of a function of X. For example, the transformation of an estimator $\hat{\theta}$ by $g(.)$, where $g(.)$ might be a quantile function. Depending on the selected transformation function $g(.)$, direct estimation of the desired distribution might range from trivial to impossible. Hence, we must rely on approximation methods to obtain the intended distribution for the latter case. The delta method provides an algorithm to obtain such approximation under certain conditions of regularity for X and $g(.)$.

Theorem 5.0.1 (Taylor's series with remainder) *Let $f \in C^{n+1}$ with domain that contains a. The Taylor formula for every x in this interval is given by*

$$f(x) = \sum_{i=1}^{n} \frac{f^i(a)}{i!}(x-a)^i + R_n(x), \tag{5.1}$$

where

$$R_n(x) = \frac{1}{n!}\int_a^x (x-1)^n f^{n+1}(t)\partial t. \tag{5.2}$$

For interested readers, [Apostol, 1967] presents a demonstration of this theorem. Regarding the remainder, it can be shown that

$$\lim_{x\to a} \frac{R_n(x)}{(x-a)^n} = 0. \tag{5.3}$$

Hence, we can approximate f by a polynomial function $P_n(x)$ with degree n. As $x \to a$, this approximation has an error of smaller order than $(x-a)^n$.

We will use a *Taylor series* of first order as a tool to obtain an approximation of $g(.)$. We assume that Z_n is a sequence of random variables such that

$$\sqrt{n}[Z_n - \theta] \xrightarrow{D} N(0, \sigma^2),\ \sigma > 0. \tag{5.4}$$

The goal is to obtain the distribution of $g(Z_n)$. Let $g(.) \in C^1$, such that $g'(z) \neq 0$. We consider the following first order Taylor expansion of $g(.)$ around θ.

$$g(Z_n) \approx g(\theta) + g'(\theta)(Z_n - \theta). \tag{5.5}$$

Note that for this approximation we assume that the remainder $R_1(Z_n)$ is 0, the reason being that we are computing the Taylor series in θ, the assumed mean value of Z_n. Hence, we infer that as long as n is large enough, $Z_n \to \theta$ in probability. Thus by (5.3), $R_1(Z_n) \to 0$ in probability.

Here we outline how to estimate the expectation and variance of $g(Z_n)$. Taking expectation on both sides of (5.3), we obtain

$$\begin{aligned} E[g(Z_n)] \approx & E[g(\theta) + g'(\theta)(Z_n - \theta)] \\ = & g(\theta) + E[g'(\theta)((Z_n - \theta)] \\ = & g(\theta) + g'(\theta)E[(Z_n - \theta)] = g(\theta). \end{aligned} \tag{5.6}$$

Similarly, we apply the variance to both sides of (5.3). Consequently,

$$\begin{aligned} Var[g(Z_n)] \approx & Var[g(\theta)+g'(\theta)(Z_n-\theta)] \\ = & Var[g'(\theta)(Z_n-\theta)] \\ = & g'(\theta)^2 Var[(Z_n-\theta)] \\ = & g'(\theta)^2\sigma^2. \end{aligned} \tag{5.7}$$

Let's "play"with the expression$\sqrt{n}[g(Z_n)-g(\theta)]$. Replacing $g(Z_n)$ by its Taylor first order approximation, we obtain

$$\begin{aligned} \sqrt{n}[g(\theta)+g'(\theta)(Z_n-\theta)-g(\theta)] = & \sqrt{n}[g'(\theta)(Z_n-\theta)] \\ = & g'(\theta)\sqrt{n}[Z_n-\theta]. \end{aligned} \tag{5.8}$$

Hence,$\sqrt{n}[g(Z_n)-g(\theta)] \approx g'(\theta)\sqrt{n}[Z-\theta]$. Using this approximation, we want to obtain the distribution family for$\sqrt{n}[g(Z_n)-g(\theta)]$. Let $B=\sqrt{n}[Z-\theta]$, $B \sim N(0,\sigma^2)$ by construction. Consequently

$$P(g'(\theta)B \leq b) = P\left(B \leq \frac{b}{g'(\theta)}\right) = F_B\left(\frac{b}{g'(\theta)}\right),$$

where $F_B(b)$ is the distribution function of B. We now calculate the derivative

$$F'_B\left(\frac{b}{g'(\theta)}\right) = f_B\left(\frac{b}{g'(\theta)}\right)\frac{d}{db}\frac{b}{g'(\theta)} = f_B\left(\frac{b}{g'(\theta)}\right)\frac{1}{g'(\theta)},$$

where

$$f_B(b) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{b^2}{2\sigma^2}}.$$

Hence

$$F'_B\left(\frac{b}{g'(\theta)}\right) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\left(\frac{b}{g'(\theta)}\right)^2}{2\sigma^2}}\frac{1}{g'(\theta)} = \frac{1}{g'(\theta)\sigma\sqrt{2\pi}}e^{-\frac{b^2}{2\sigma^2 g'(\theta)^2}},$$

which corresponds to the density function of Normal distribution $N(0,\sigma^2 g'(\theta)^2)$. Thus, $g'(\theta)\sqrt{n}[Z-\theta] \approx N(0,\sigma^2 g'(\theta)^2)$. Since$\sqrt{n}[g(Z_n)-g(\theta)] \approx g'(\theta)\sqrt{n}[Z-\theta]$, we say that$\sqrt{n}[g(Z_n)-g(\theta)] \sim N(0,\sigma^2 g'(\theta)^2)$.

This methodology is well suited to obtain the distribution of transformations of maximum-likelihood estimators. We will now present the case where we intend to infer on a transformation of a vector of maximum likelihood estimators. Let $\hat{\theta}=(\hat{\theta_1},\hat{\theta_2},...,\hat{\theta_m})$, the vector of estimators with means $E(\hat{\theta_i})=\theta$ such that $\hat{\theta} \xrightarrow{D} N(0,\Sigma)$, where Σ is the variance-covariance matrix of $\hat{\theta}$. Let $g:\mathbb{R}^n \longrightarrow \mathbb{R}$ with non null first order partial derivatives. The Taylor's first order expansion of $g(.)$ around $\hat{\theta}=(\theta_1,\theta_2,...,\theta_m)$ is as follows:

$$g(\hat{\theta}) = g(\theta)+\nabla^T g(\theta)(\hat{\theta}-\theta), \tag{5.9}$$

where $\nabla^T g(\theta)$ is the gradient vector of $g(.)$, $\nabla^T g(\theta)=[\frac{\partial g}{\partial\theta_1},\frac{\partial g}{\partial\theta_2},...,\frac{\partial g}{\partial\theta_m}]$.Taking expectations on both sides, we obtain

$$\begin{aligned} E[g(\hat{\theta})] = & E[g(\theta)+\nabla^T g(\theta)(\hat{\theta}-\theta)] \\ = & g(\theta)+\nabla^T g(\theta)(E[\hat{\theta}]-\theta) \\ = & g(\theta). \end{aligned} \tag{5.10}$$

We now apply the variance to both sides of (5.9), hence obtaining

$$\begin{aligned} Var[g(\hat{\theta})] = & Var[g(\theta)+\nabla^T g(\theta)(\hat{\theta}-\theta)], \\ = & Var[\nabla^T g(\theta)(\hat{\theta}-\theta)] \\ = & \nabla^T g(\theta)\Sigma\nabla g(\theta). \end{aligned} \tag{5.11}$$

Similar to the univariate case, we can show that $\sqrt{n}[g(\hat{\theta}) - g(\theta)] \xrightarrow{D} N(0, \nabla^T g(\theta) \Sigma \nabla g(\theta))$. Confidence intervals for the univariate and multivariate cases can be constructed using the following formulas:

$$IC_{1-\alpha} = g(\theta^*) \pm z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{g'(\theta^*)^2 \, \sigma^2}{n}}, \tag{5.12}$$

$$IC_{1-\alpha} = g(\theta^*) \pm z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\nabla^T g(\theta^*) \Sigma \nabla g(\theta^*)}{n}}, \tag{5.13}$$

where $z_{1-\frac{\alpha}{2}}$ is the quantile with probability $1 - \frac{\alpha}{2}$ from a standardized normal distribution, α is the selected significance level and θ^* is the maximum likelihood estimate of θ.

6 | Descriptive Analysis of the Biometric Variables Recorded in Portuguese Voluntary Pharmacy Attendees

6.1 Introduction

Hypertension, also known as high blood pressure, is described as an abnormal pressure on the blood vessels caused by blood flow. As blood is pumped through the body, the blood vessels are impacted by this flow, thus creating blood pressure and blood vessel tension. The higher the tension, the more strength the heart needs in order to pump the blood. Diagnosing hypertension is done by measuring two blood pressure markers. Systolic blood pressure is the tension measured by the compliance of the blood vessels to the blood flow during a heartbeat. Diastolic blood pressure is the tension measured between heartbeats.

According to the World Health Organization (WHO), hypertension is a global public health issue. It is highly associated with incidents of heart disease, stroke, kidney failure, premature mortality and disability. In Portugal medical practitioners state that half of the adults of age 40 and higher suffer from hypertension and one third has not been diagnosed. This is due to most of the early onset symptoms of the disease being very mild. It is also the risk factor most associated with the leading death causes in the country.

Hypertension has been linked to unhealthy diets, sedentary lifestyle, drug abuse and tobacco use, see [Schröder et al., 2003] and [Wakabayashi, 2004]. The former studies the relationship between diet, body mass index, cholesterol, leisure-time physical activity and diet on the Mediterranean Southern-Europe population, while the latter studies the relationship of the body mass index with blood pressure and serum cholesterol concentrations at different ages. See also [Hajar, 2016], where the author outlines the history of the study of risks associated with hypertension.

With the goal of addressing this public health issue, the Portuguese National Association of Pharmacies developed a campaign in 2005 through their Department of Pharmaceutical Care to study the risk factors associated with the leading death causes in the country.

In the following section, we will descriptively analyze several biometric variables as a factor of some other measures of interest (on the individuals who suffer from isolated systolic hypertension), such as the relation between systolic blood pressure and age, tobacco consumption, body mass index and gender.

In the next two chapters, we will apply the aforementioned *Peaks Over Threshold* methodology to the individuals who suffer from isolated systolic hypertension, which are characterized by having **diastolic blood pressure** < 90 **mmHg and systolic pressure** ≥ 140 **mmHg**. The classification categories in terms of blood pressure conditions are presented in table 6.1 (guidelines of the Portuguese Cardiology Association). The goal is to fit a generalized Pareto distribution to the excesses above a high threshold u for each Portuguese district, and subsequently estimate tail probabilities and extreme quantiles. With this study, we hope to answer questions about the condition of the most extreme isolated systolic hypertension cases for each district.

Table 6.1: Categories of blood pressure in mmHg (Portuguese Cardiology Association guidelines)

Category	**Systolic Blood Pressure**	**Diastolic Blood Pressure**
Optimal	< 120	< 80
Normal	120-129	80-84
Normal High	130-139	85-89
First Degree Hypertension	140-159	90-99
Second Degree Hypertension	160-179	100-109
Third Degree Hypertension	≥ 180	≥ 110
Isolated Systolic Hypertension	≥ 140	< 90

6.2 Exploratory Data Analysis

In this section, we aim to create a descriptive profile of the individuals from different cohorts of the hypertension pathology. There are a total of 40065 individuals in this database and 34 variables. To that effect, we need to split the individuals into four distinct cases. The healthy individuals, who have systolic blood pressure (SBP) less than or equal to 90 mmHg and diastolic blood pressure (DBP) less than 140 mmHg ($n = 18113$). The rarer case of individuals who have readings of DBP higher than 90 mmHg and readings of SBP lower than 140 mmHg ($n = 1076$). The individuals that have both DBP and SBP higher than the standard thresholds ($n = 7500$) and the individuals who suffer from the aforementioned isolated systolic hypertension ($n = 9996$). The latter consists makes up the bulk of hypertension pathology, thus being the focus of this dissertation. There are also 3380 individuals with omitted information about these variables.

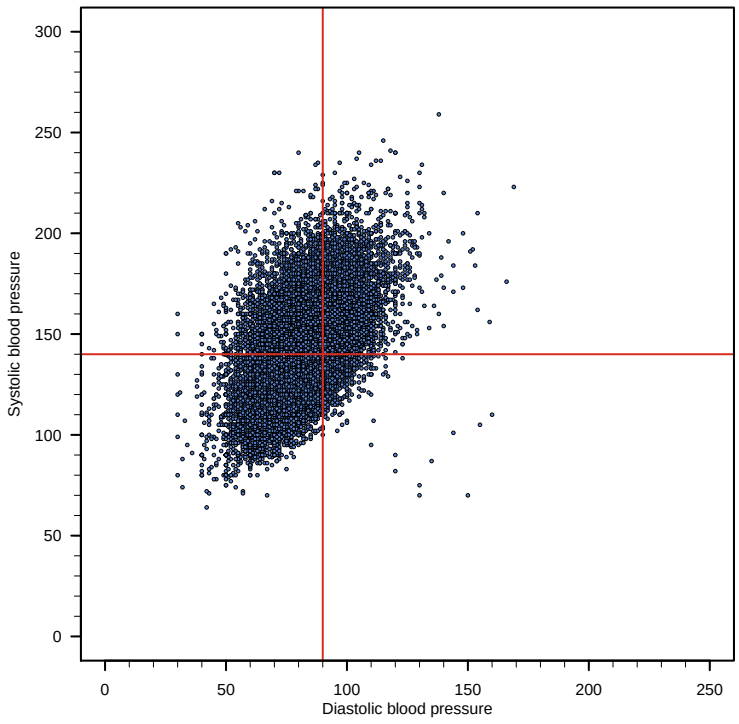

Fig. 6.1: Diastolic blood pressure *vs*. systolic blood pressure for Portuguese voluntary pharmacy attendees

Figure 6.1 is the result of plotting diastolic blood pressure versus systolic blood pressure for Por-

tuguese voluntary pharmacy attendees, giving rise to the aforementioned stratification. It also suggests some linear correlation with positive slope between these two blood markers. The Pearson's correlation coefficient for the whole dataset yields the value 0.5643, which supports the previous statement. It would be interesting to study possible correlations between the extreme values of both variables in the sphere of extreme value theory. Though a considerable amount of literature exists regarding bivariate extreme value analysis, this topic falls outside the scope of this dissertation. The red horizontal and vertical lines convey the accepted limits over which an individual is considered as suffering from an hypertension-type pathology, as illustrated by table 6.1.

For the following analysis, we consider the individuals who suffer from isolated systolic hypertension. The focus on this pathology is due to the fact that it is the most prevalent form of hypertension in this dataset.

Table 6.2: Summary statistics of the systolic blood pressure in men and women who suffer from isolated systolic hypertension

Gender	**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**Prop**
Male	140.0	144.0	150.0	153.3	160.0	230.0	0.3792
Female	140.0	144.0	150.0	154.1	160.0	235.0	0.6192

Table 6.2 illustrates the behavior of the SBP in men and women who suffer from isolated systolic hypertension. Out of the 9996 individuals, approximately 62% are women and 38% are men. There are 18 observations with missing gender values. Also note that this population has a lower SBP bound of 140 mmHg. There isn't a clear difference between the two groups regarding the SBP mean. Furthermore, the female group has a higher 3rd quartile than the male group and also a larger maximum value, which suggests that men and women may have some SBP differences in the extremes.

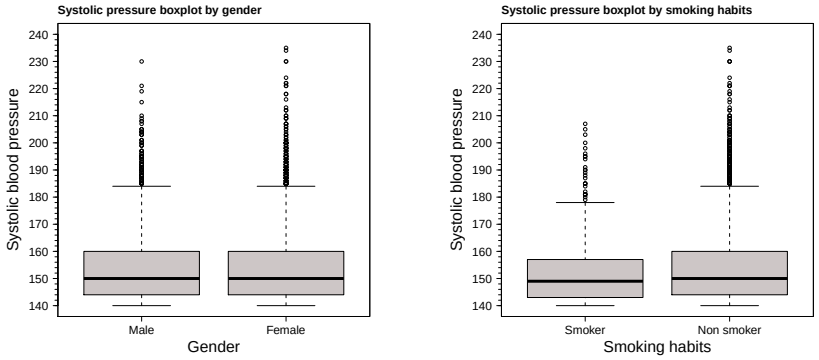

Fig. 6.2: Systolic blood pressure boxplots by gender (left) and by tobacco consumption (right)

Figure 6.2 illustrates the SBP boxplots by gender and tobacco consumption. It can be seen that women seem to have more and higher extreme values of SBP than men. As mentioned before, there is some literature on the association between tobacco consumption and high values of blood pressure. In this case, the boxplot produced seems to indicate that those who smoke seem to have overall lower values of SBP than those who don't. However, no credible conclusion about this relation can be derived from this boxplot, since there are many confounding factors. For instance, out of the 9586 individuals with recorded smoking habits, only 6.3% are smokers. Moreover, this boxplot includes men and women, young and old individuals, which might influence this outcome. Also, to the best of our knowledge, there was no verification of the veracity of the attendees' answers. One could simply lie about his or her tobacco smoking habits.

One variable that has been shown to be highly associated with high values of systolic blood pressure is age, see [Pinto, 2007] and [Bavishi et al., 2016].

Table 6.3: Summary of the systolic pressure by age in Portuguese voluntary pharmacy attendees who suffer from isolated systolic hypertension

Age	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Prop
<25	140.0	140.8	145.0	149.2	150.5	206.0	0.0059
25-34	140.0	141.0	146.0	149.2	152.0	212.0	0.0180
35-44	140.0	142.0	147.0	149.2	152.0	207.0	0.0331
45-54	140.0	142.0	148.0	149.7	154.0	203.0	0.0844
55-64	140.0	143.0	149.0	152.0	158.0	213.0	0.2180
65-74	140.0	144.0	150.0	154.1	160.0	240.0	0.3796
>=75	140.0	145.0	154.0	157.2	165.0	235.0	0.2611

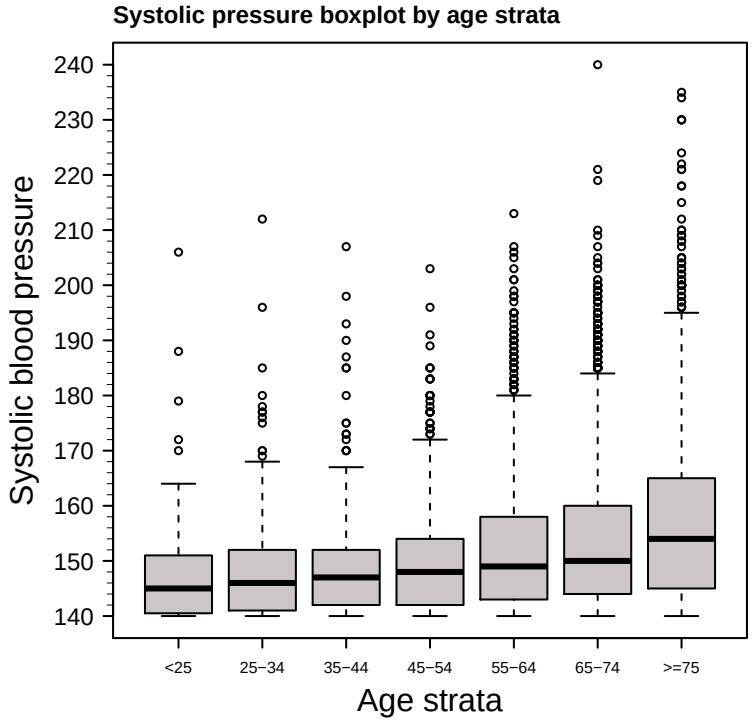

Fig. 6.3: Systolic blood pressure boxplots by age strata

Table 6.3 presents a summary of the SBP variable in an array of different age strata. The individuals are not equally distributed by age stratum. The bulk of the observations lie above the 55 age group. This might be the result of selecting individuals who suffer from isolated systolic hypertension. This pathology is known to be more common in the elderly.

We can see a steady rise of overall SBP values as age goes up. This can also be observed in figure 6.3, where it is most apparent that older individuals tend to have overall higher values of SBP. SBP is the tension the blood flow produces on the blood vessels during a heartbeat. As a person gets older, they tend to lose blood vessel elasticity, thus increasing the tension generated by the blood flow.

The body mass index (BMI) is a value used to ascertain the amount of fat tissue in an individual. It is expressed in kg/m^2 and is given by

$$BMI = \frac{BodyMass}{Height^2}.$$

Table 6.4 illustrates the different classes of body mass index. Our interest is to study the possible relation between high values of systolic blood pressure and high values of BMI. To that effect, we obtained the boxplots for the individuals in each of these categories and their respective summary statistics. Table 6.5 and figure 6.4 illustrate the results.

It is important to notice that, similar to the age variable, the individuals are not evenly spread amongst the different BMI strata, since the bulk of the data is constituted by overweight and obese individuals (more than 80%). We remind the reader that these are individuals that suffer from isolated systolic hypertension, and the uneven size distribution in the strata might therefore be a consequence of this fact; maybe the prevalence of this pathology is higher in individuals with high BMI. The low number of observations in the category of underweight individuals might also be due to the fact that underweight people tend to have lower values of systolic blood pressure, hence underweight individuals exceeding 140 mmHg are rare. Not taking into account the underweight stratum, we can observe that there is little to no difference in blood pressure levels between each class. This suggests that BMI by itself may not be sufficient to account for high levels of systolic blood pressure.

Table 6.4: BMI classes

Underweight	**Normal Weight**	**Overweight**	**Obese**
<18.5	18.5-24.9	25-29.9	>30.0

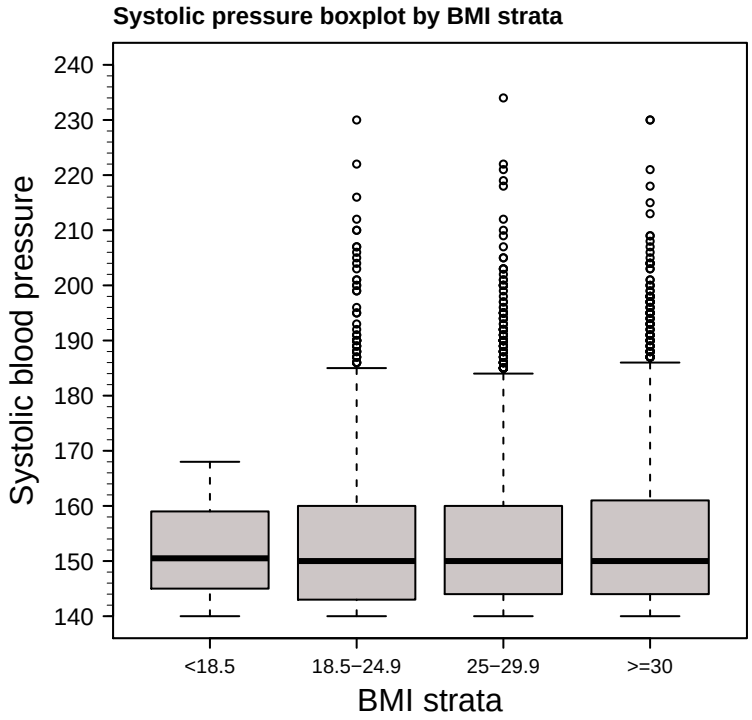

Fig. 6.4: Systolic pressure by BMI strata

Our next study interest is to compare the individuals relative to their systolic blood pressure,in each Portuguese district and islands. There are several goals we want to achieve with this analysis. First, we want to compare rural districts, usually characterized by having low population density, with metropolitan districts, which are frequently associated with high population density. The objective is to compare a more nature-bound lifestyle with the health conditions related to living in a city, since, as stated before, stress and a sedentary lifestyle are considered risk factors for cardiovascular diseases. These are strongly

BMI stratum	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Prop
<18.5	140.0	145.2	150.5	151.4	158.2	168.0	0.0027
18.5-24.9	140.0	143.0	150.0	153.3	160.0	230.0	0.1918
25-29.9	140.0	144.0	150.0	153.3	160.0	234.0	0.4718
>30	140.0	144.0	150.0	154.5	161.0	230.0	0.3337

Table 6.5: Summary of the systolic blood pressure by BMI strata in Portuguese voluntary pharmacy attendees who suffer from isolated systolic hypertension

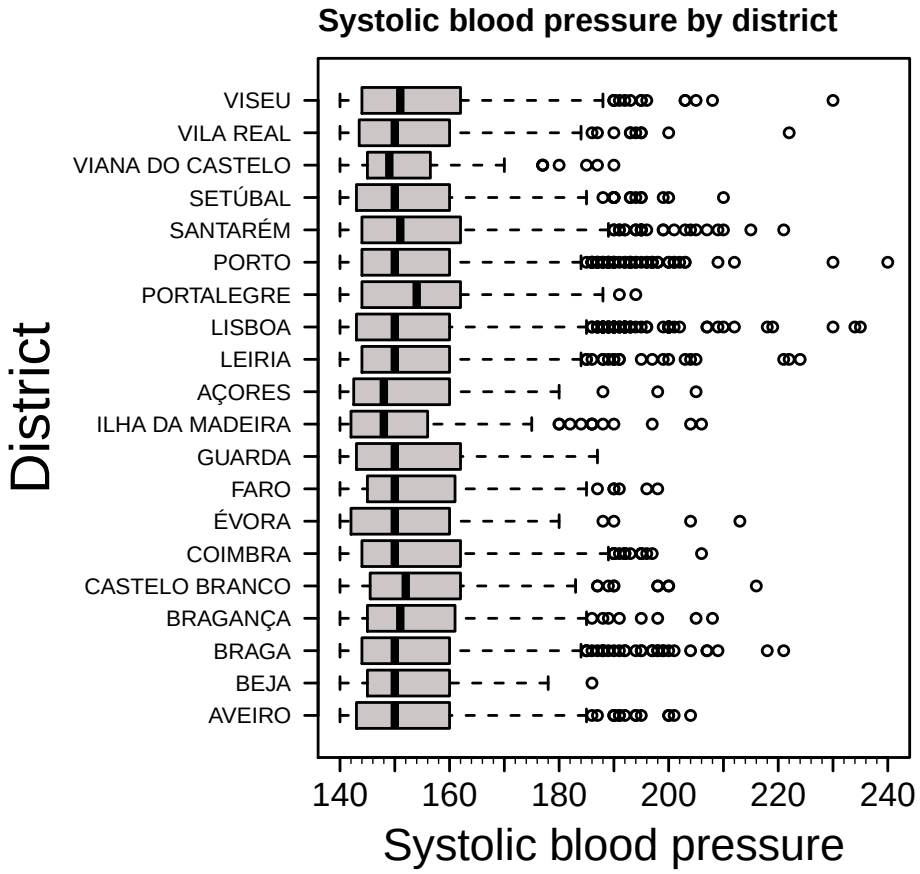

Fig. 6.5: Systolic blood pressure by Portuguese district

associated with high values of systolic blood pressure. Secondly, we want to map out the systolic blood pressure profile for individuals suffering from isolated systolic hypertension, so it can serve as a tool for medical services. Figure 6.5 illustrates the boxplots of the values of systolic blood pressure observed in individuals who suffer from isolated systolic blood pressure by Portuguese district and islands. One curious phenomena, illustrated in table 6.6, is that higher population density districts yield higher maximum values, i.e., the largest value is observed in Porto, which is the second most populated Portuguese district, followed by Lisboa, the most populated district, with a maximum equal to 235 mmHg. Some other high population density districts with extreme maximum values are: Braga, with maximum 229 mmHg, which is the third most populated Portuguese district and Viseu, with maximum 230 mmHg. The previously mentioned districts were also the ones that supplied the largest samples (with the exception of Viseu), which might also be the cause for such high maximum values when compared to other districts with smaller sample size. Regarding the median values of systolic blood pressure, the islands Açores and

Table 6.6: Summary statistics of systolic blood pressure by Portuguese district and islands

District	Min	1st Qu.	Median	Mean	3rd Qu.	Max	n
Aveiro	140.0	143.0	150.0	152.8	160.0	204.0	736
Beja	140.0	145.0	150.0	153.2	160.0	186.0	117
Braga	140.0	144.0	150.0	154.7	160.0	221.0	810
Bragança	140.0	145.0	152.0	154.7	161.0	208.0	276
Castelo Branco	140.0	146.0	153.0	156.4	163.0	216.0	200
Coimbra	140.0	144.0	150.0	154.5	162.0	206.0	441
Évora	140.0	144.0	150.0	153.2	160.0	213.0	186
Faro	140.0	145.0	150.0	154.1	160.0	198.0	227
Guarda	140.0	143.0	150.0	154.1	160.2	187.0	103
Ilha da Madeira	140.0	140.0	147.0	150.7	155.0	206.0	221
Açores	140.0	142.8	149.0	152.3	160.0	205.0	88
Leiria	140.0	144.0	150.0	154.4	161.0	224.0	467
Lisboa	140.0	143.0	150.0	153.5	160.0	235.0	2248
Portalegre	140.0	145.0	154.0	154.9	162.0	194.0	92
Porto	140.0	144.0	150.0	154.2	160.0	240.0	1590
Santarém	140.0	144.0	151.0	156.1	164.0	221.0	550
Setúbal	140.0	143.0	150.0	152.7	160.0	210.0	788
Viana do Castelo	140.0	145.0	150.0	152.9	160.0	190.0	120
Vila Real	140.0	144.0	150.0	153.7	160.0	222.0	299
Viseu	140.0	144.0	151.0	155.8	162.0	230.0	320

Madeira, which are the only non-continental Portuguese districts, provided lower median systolic blood pressure than any individual district from the mainland. The remaining districts are similar, with the exception of Portalegre, which has a slightly higher median than the rest. We mention again the article [Schröder et al., 2003], where the authors study the relationship between diet, leisure activity, BMI and serum cholesterol. Such an analysis would also be well suited for the Portuguese districts and islands, since an individual's diet and lifestyle varies geographically.

7 | First Approach to Extreme Value Modeling of Systolic Blood Pressure Values

7.1 Data Description and Methodologies

As stated before, the aim of this dissertation is to apply methodologies to build models for extreme values of systolic blood pressure in individuals who suffer from isolated systolic hypertension. One could start by developing a model for the complete dataset. We've decided not to, since the exploratory data revealed that there may be several confounding factors that might make this study extremely hard. Later in this chapter we will discuss several of the difficulties associated with this analysis. In this section, we propose a preliminary model using the previously mentioned *Peaks Over Threshold* methodology for the district of Braga. The aim is to provide a model for SBP extremes for this district and to illustrate the methods used, as well as the difficulty in their implementation. With the goal of not being too repetitive, we will present the complete analysis for Braga and then only present the final results for the remaining districts.

Our dataset consists of voluntary individuals who attended one of the pharmacies of the district of Braga that joined the campaign, and who suffered from isolated systolic hypertension, i.e., individuals with SBP higher than or equal to 140 mmHg and DBP lower than 90 mmHg. The previous chapter covers the exploratory analysis of these individuals.

The objective is to fit a model to the excesses over a high value u using this data. In this case, a generalized Pareto distribution with shape parameter $k \in \mathbb{R}$ and scale parameter $\sigma > 0$

$$F(x) = \begin{cases} 1 - \left(1 + \frac{kx}{\sigma}\right)^{-\frac{1}{k}} & k \in \mathbb{R} \setminus \{0\}, \\ 1 - e^{-\frac{x}{\sigma}} & k = 0, \end{cases} \tag{7.1}$$

with support $\left\{x \in \mathbb{R} : x > 0\right\}$ for $k \geq 0$ and support $\left\{x \in \mathbb{R} : 0 < x < -\frac{\sigma}{k}\right\}$ for $k < 0$.

Before we can proceed with the estimation of the GPD parameters via the maximum likelihood method, we must select the threshold u. We've presented several methodologies in a previous chapter to choose u. [DuMouchel, 1983] claims that a simple rule is to consider the $\chi_{0.90}$ sample quantile. This choice would yield a sample of approximately 91 individuals for Braga. This is a reasonable sample size to apply ML estimation, but it is not ideal. Moreover, the application of this rule to some of the districts would not be feasible due to the reduced size of the sample of excesses. For example, if we intend to build a model for the district of Açores by selecting the 90% sample quantile, we would end up by having only 7 excesses, which is clearly not sufficient to fit a reasonable asymptotic model. Another issue is that this dataset has a lower bound. Meaning, we selected individuals with systolic blood pressure values above 140, which might suggest that for some districts we might already be in the tail of the distribution. Nevertheless, higher threshold values should still provide an adequate data fit to the GPD model.

We now introduce the procedure implemented in this dissertation as a first approach to extreme value analysis. We selected Braga as an example and will explain each step of this analysis in detail. Later, we will present the results for the remaining districts using the same procedures.

7.2 Model Fitting

We begin by considering an array of threshold candidates $(140, 150, 160, 170, 180)$ for Braga, for $n = 810$ individuals. We expect that one of these values should be an adequate threshold. We didn't consider values above 180 mmHg since there are only 41 observations higher than this value in this district.

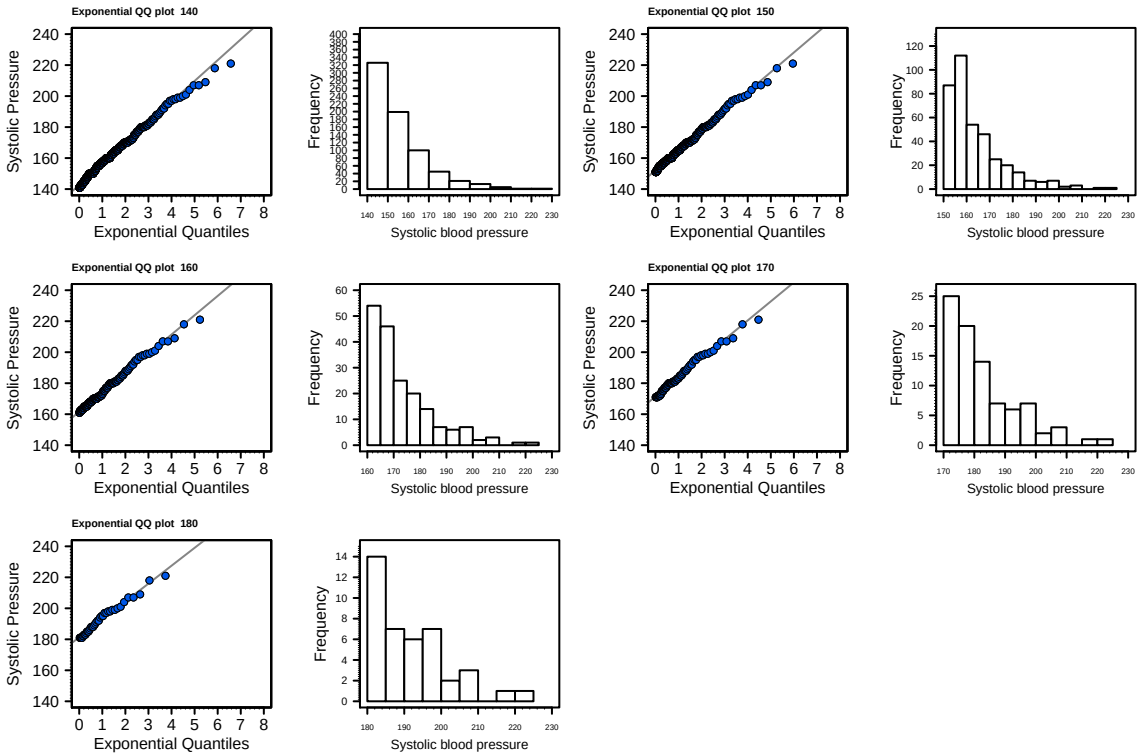

Fig. 7.1: Exponential QQ-plots and histograms for an array of thresholds for Braga

We now consider the five datasets consisting of the values above each threshold candidate. Figure 7.1 corresponds to the exponential QQ-plots and the histograms for each dataset. The exponential QQ-plots try to measure the compatibility of the given dataset with the exponential distribution by comparing the empirical quantiles with the exponential quantiles, by plotting the latter *vs.* the former. We expect the result to reflect a linear dependency. The histograms consist in plotting the dataset's raw distribution. We can see that all the datasets seem to display a reasonable fit to the model. The QQ-plots also suggest that the distribution has an exponential tail.

Figure 7.2 contains the plot of the mean residual life function. The sampled mean residual life function is given by

$$\hat{e}(u) = \frac{1}{n_u} \sum_{i=1}^{n_u} \left(x_{i_u} - u \right), \quad u < \max(x_1, x_2, ..., x_n), \tag{7.2}$$

where x_{i_u} are the exceedances over u. As presented in (7.2), we plot a range of thresholds versus the mean of the excesses above each threshold. As mentioned before, we expect to find a linear behavior above some high value for u. This plot is very hard to interpret. The expected linear-like behavior cannot

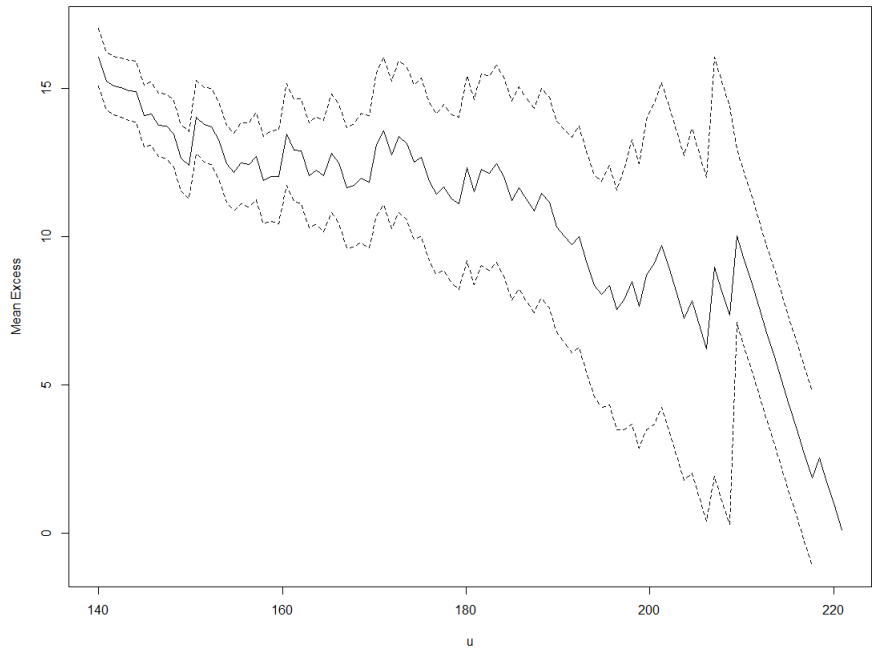

Fig. 7.2: Estimated mean residual life function for the Braga district

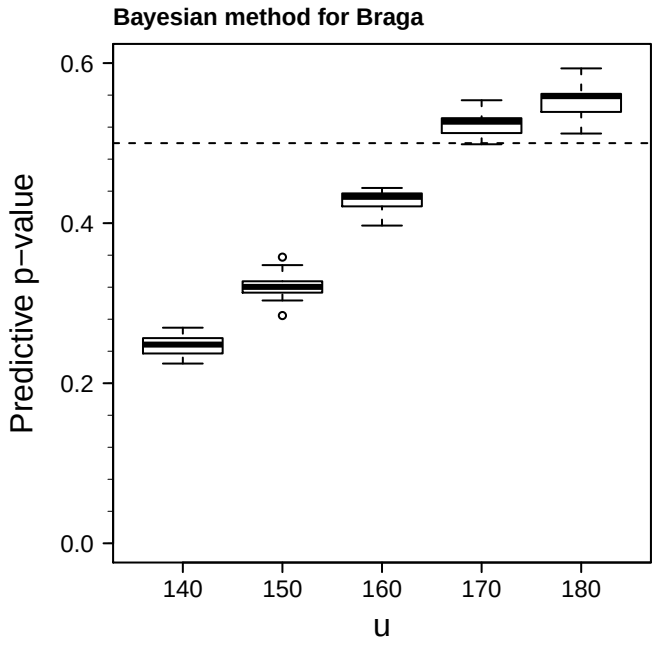

Fig. 7.3: Bayesian method for threshold selection using measure of surprise

be construed from this plot at any given partition of its domain. The method's output was obtained by using the *mrl.plot* function from the R package *ismev*.

Figure 7.3 presents the results from the Bayesian method for threshold selection mentioned in chapter 3. The results seem to indicate that as we *climb the threshold ladder*, there is less data incompatibility with the GPD model. It suggests that for a threshold value between 160 mmHg and 170 mmHg, the predictive p-values appear to be sufficiently close to 0.5, which is the expected predictive p-value under the null hypothesis of the GPD fit. This method was applied with the aid of R code provided by the authors, see [Lee et al., 2015].

Next, we consider the following hypotheses. Let u_i, $i = 1, ..., m$, be a candidate threshold where

Table 7.1: Cramér-von Mises and Anderson-Darling hypothesis testing for the Braga district

u	$\hat{k}$	W^2	A^2	p_{A^2}	p_{W^2}
140	-0.1747	0.7144	5.0807	~ 0	~ 0
150	-0.1609	0.6048	4.2326	~ 0	~ 0
160	-0.1681	0.2536	2.0084	~ 0.001	~ 0
170	-0.2162	0.0608	0.4814	~ 0.25	~ 0.25
180	-0.2921	0.0318	0.2901	~ 0.5	~ 0.5

$u_1 = 140$ mmHg and $u_5 = 180$ mmHg. Let's consider the following sequential tests of hypotheses: H_0^i : the observed data $z_{n_i} = \{z \in S : z > u_i\}$ comes from a generalized Pareto distribution, where S is the sample for the district of Braga and $i = 1, 2, ..., m$. Sequential testing should be handled with care, since if we reject the kth hypothesis, we must have rejected all that preceded it. Note also that in this chapter we do not yet account for multiple testing. This issue will be discussed in a future chapter using rules, as addressed by [Bader et al., 2018], to minimize the risk of falsely selecting a low threshold. Hence, we decided to perform the tests on 5 candidates. For each hypothesis test we compute two test statistics: the Anderson-Darling and the Cramér-von Mises statistics for the GPD. Using the algorithm outlined in previous chapters, we obtain the p-values. Next, we present the test statistics for the referred tests

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\big[\ln(z_i) + \ln(1-z_{n+1-i})\big], \quad W^2 = \sum_{i=1}^{n}\big[z_i - \frac{(2i-1)}{(2n)}\big]^2 + \frac{1}{12n}.$$

It has been shown that for large samples, these tests have incredible accuracy rates, as described in [Bader et al., 2018], where the authors perform several accuracy simulations for each test using known distributions. Table 7.1 illustrates the resulting test statistics and p-values for each candidate. There is a sudden raise in the p-value from 160 mmHg to 170 mmHg, leading to the rejection of the hypothesis at 160 mmHg and non rejection at 170 mmHg, at the significance level of 0.05, which might indicate that an adequate threshold lies between these two values. Note also that an estimate of k is also included, since the test statistics of both tests are parameter-dependent. See [Choulakian and Stephens, 2001] for the distribution tables for each test. These outputs were obtained by programming R functions for each test.

It is remarkable, not accounting for small differences, that all the methods seem to point towards the same threshold range. This evidence is also backed up by the exploratory analysis of the exponential QQ-plots and histograms. The threshold selected was 164 mmHg. The reason behind the choice of this value is because it was the first value for which we did not reject the null hypothesis for each test.

Table 7.2: Model fitted to Braga

u	n	$\hat{k}$	95% CI for k	$\hat{\sigma}$	95% CI for σ	max	endpoint	$-\ln(L)$
164	150	-0.096	(-0.265;0.074)	13.79	(10.58;16.00)	221	308.24	529.22

Table 7.2 pertains to the point and interval estimation of the GPD parameters via maximum likelihood. These results were obtained using the *gpd.fit* function from the R package *evir*. The obtained estimate of k is negative which suggests that high systolic blood pressure values in these individuals has a light-tailed distribution. The 95% confidence interval for k includes 0, which suggests that perhaps there is no difference between considering a GPD model or an exponential model, though this interval is severely skewed to the negative side of the axis. We remind the reader that a GPD with $k = 0$ degenerates into an exponential distribution. The Occam's razor principle, also known as parsimony, states that it is best to select the least complicated model to work with as long as there is no significant statistical

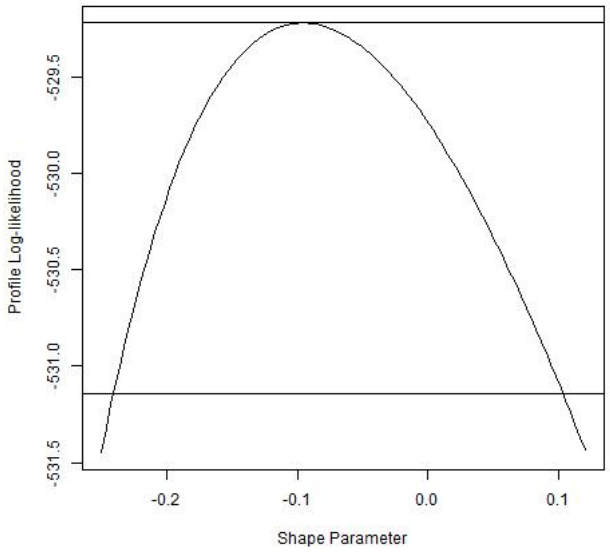

Fig. 7.4: Profile likelihood function for the shape parameter

difference between the two. Hence, we intend to test if the exponential model is sufficient. This conjecture could also be backed up by figure 7.4, which consists in plotting the profile-likelihood function for $\hat{k}$, which according to [Coles, 2001] might produce better confidence intervals for the parameters. It appears though that the curve depicted is also slightly skewed to the negative side of the axis. Still, it is worthwhile to proceed with hypothesis testing.

Our hypotheses consist of $H_0: k = 0$ *vs* $H_1: k \neq 0$. We use the deviance function as the test statistic. We can easily show that

$$T = 2\big(l_{M_1}(\mathbf{x}) - l_{M_2}(\mathbf{x})\big) \sim \chi^2_1,$$

where in this case $l_{M_1}(\mathbf{x})$ is the log-likelihood function for the GPD model and $l_{M_2}(\mathbf{x})$ is the log-likelihood function for the exponential model, and $\mathbf{x}$ is the sample used to obtain the estimates of the GPD parameters. R functions were implemented to obtain these results.

Table 7.3: Result of the deviance test for Braga

$l_{M_1}(\mathbf{x})$	$l_{M_2}(\mathbf{x})$	T	p
-529.2197	-530.3549	2.270363	0.131869

Table 7.3 illustrates the results of this test. The test yielded a high p-value and, thus, the null hypothesis should not be rejected at the usual significance levels, suggesting that the exponential model should be sufficient to explain the high values of blood pressure in the considered population.

We now present models for each district using the procedure previously outlined for Braga. The models were fitted using the *gpd.fit* function from the R package *evir*. Table 7.4 depicts the resulting extreme models for each district. We would like to point out that all the districts produced negative shape parameters, thus indicating that systolic blood pressure in individuals with isolated systolic hypertension is a phenomena whose distribution is light-tailed. Additionally, it is obvious from a biological point of view that the SBP has a finite upper bound. This characteristic of the GPD is only observed in light-tailed situations. For the models that yielded a negative shape parameter significantly close to 0, an exponential distribution was selected (just as the case with the district of Braga). Different significance levels were considered for both Anderson-Darling and Cramér-von Mises so some consistency between the methods

Table 7.4: Fitted GPD models to each Portuguese district and islands

District	n	$n > u$	$\hat{k}$	95% CI for k	$\hat{\sigma}$	95% CI for σ	Max	u	Model	A^2	W^2	α
Açores	88	73	−0.114	(-0.339,0.112)	16.76	(11.38,22.13)	205	140	EXP	0.418	0.050	0.05
Aveiro	736	154	−0.337	(-0.467,-0.208)	17.18	(13.79,20.57)	204	160	GPD	0.561	0.096	0.01
Beja	117	31	−0.380	(-0.630,-0.129)	11.84	(7.03,16.65)	186	159	GPD	0.723	0.116	0.05
Braga	810	150	−0.096	(-0.265,0.074)	13.79	(10.58,17.00)	221	164	EXP	0.939	0.122	0.01
Bragança	276	136	−0.179	(-0.324,-0.033)	15.89	(12.39,19.40)	208	151	GPD	0.686	0.087	0.05
Castelo Branco	200	40	−0.081	(-0.415,0.253)	12.33	(6.72,17.94)	216	168	EXP	0.649	0.098	0.05
Coimbra	441	209	−0.326	(-0.441,-0.212)	20.78	(17.20,24.35)	206	150	GPD	0.775	0.089	0.05
Évora	186	30	−0.023	(-0.371,0.324)	12.52	(6.28,18.76)	213	160	EXP	1.067	0.191	0.01
Faro	227	189	−0.269	(-0.392,-0.145)	19.06	(15.54,22.57)	198	142	GPD	0.758	0.081	0.05
Guarda	103	92	−0.418	(-0.627,-0.209)	22.79	(16.41,29.16)	187	140	GPD	0.453	0.067	0.05
Ilha da Madeira	221	77	−0.184	(-0.395,0.027)	17.30	(12.03,22.56)	206	150	GPD	0.677	0.104	0.05
Leiria	467	201	−0.142	(-0.254,-0.030)	18.04	(14.85,21.23)	224	151	GPD	1.275	0.174	0.01
Lisboa	2248	92	−0.017	(-0.215,0.182)	11.44	(8.18,14.70)	235	181	EXP	0.916	0.122	0.05
Portalegre	92	86	−0.300	(-0.485,-0.115)	20.51	(14.91,26.12)	194	140	GPD	0.515	0.070	0.05
Porto	1590	203	−0.034	(-0.154,0.085)	11.30	(9.24,13.36)	240	169	EXP	1.396	0.154	0.01
Santarém	550	39	−0.407	(-0.670,-0.144)	19.07	(11.65,26.49)	221	180	GPD	0.335	0.047	0.05
Setúbal	788	169	−0.235	(-0.344,-0.125)	14.59	(11.92,17.27)	210	160	GPD	1.166	0.139	0.01
Viana do Castelo	120	105	−0.199	(-0.375,-0.024)	15.16	(11.26,19.05)	190	141	GPD	1.277	0.197	0.01
Vila Real	299	158	−0.078	(-0.220,0.066)	14.29	(11.27,17.31)	222	149	EXP	0.919	0.117	0.05
Viseu	320	87	−0.149	(-0.324,0.025)	17.43	(12.70,22.16)	230	160	EXP	0.558	0.081	0.05

could be achieved. Some threshold choices could be considered controversial, for example, Beja has only 31 exceedances, which might just be low enough to violate the asymptotic properties of the model. We will highlight the difficulties of this analysis at the end of this chapter.

Lastly, we present the predicting capabilities of the models for each district. Table 7.5 presents extreme empirical quantiles and extreme model quantiles for each district's GPD model. These were computed using the *tailplot* function from the *evir* R package. Table 7.6 displays the quantile estimates for the districts where the exponential model was deemed suitable. These were obtained using the asymptotic properties of the ML estimator of σ and the delta method. The goal is to compare how the models hold up to the data. It is important to note that comparison does not serve as a true accuracy measure of the model's performance since it is biased to estimate extreme empirical quantiles with little observations. It is expected that in the cases where there is an abundance of observations, the models have sufficient prediction accuracy. Confidence intervals for the quantiles were obtained through the profile likelihood function.

7.3 Difficulties of a First Approach to Extreme Value Analysis

In this section we present the adversities and complications associated with this analysis. Most of these are common with other statistical approaches, but we feel it is worthwhile to discuss them.

The database consists of a sample of 40065 individuals with 34 observed variables for each individual. Of these, 9996 were used in this analysis (actually 9879 were used since there were 117 missing district values). For the sake of simplifying the analysis, all of the incomplete or missing values from variables of interest were not considered.

We start with the quantized structure of the data. The data is formed only by integers, meaning only integer observations were recorded. This is due to the nature of the procedure used to obtain these values.

Figure 7.5 illustrates one of the main issues with the observations in this dataset. *Rounded* numbers like 140, 150, 160, 170,..., 200 have much higher frequencies than their neighbors. This was probably due to biased approximations or machinery that provided less precise readings. This issue was not apparent in the first exploratory steps of the data analysis. It became evident much later when applying methods for threshold selection. It appeared that on these *rounded* numbers the test statistics for Anderson-Darling and Cramér-von Mises tests would suddenly drop. Moreover the Bayesian method demonstrated that perhaps there was more than one candidate threshold for the data, possibly implying a mixture of distributions.

We now elaborate more on this issue using the district of Coimbra, which was the hardest to analyze due to being difficult to find common ground between all the threshold selection methods.

Figure 7.6 illustrates the estimated kernel density function for the SBP values of individuals from the district of Coimbra who suffer from ISH. The issue of higher frequencies for *round* numbers persists. We now illustrate how this might affect some of the methodologies for threshold selection.

We point out the differences between the graphics of Bayesian method for Braga figure 7.3 and Coimbra 7.7. Note that we decided to compute the Bayesian method for Coimbra using the following threshold candidates (140,145,150,155,160,165,170,175,180), in order to better illustrate this issue. Braga has a regular behavior, since it is expected that, for higher thresholds, there is less data incompatibility with the GPD model. As stated before, if a GPD can be fitted for a threshold u_0, it can also be adequately fitted for u_1, $u_1 > u_0$, and they share the same shape parameter. In the other hand, Coimbra's application of the method suggests that thresholds above 165 show less compatibility with the GPD model. The difference might be explained by the more continuous nature of the district of Braga as illustrated by figure 7.8.

It is important to state that despite this issue, the resulting models for each district still perform well. We will address measures to attenuate this issue in the next chapter, where we will *shake* the sample in order to obtain a more continuous and smoother distribution without corrupting the data.

Now we revisit the threshold selection method using hypothesis testing. Specifically, the Anderson-Darling and Cramér-von Mises tests for the GPD. Table 7.7 illustrates the results of these tests for the

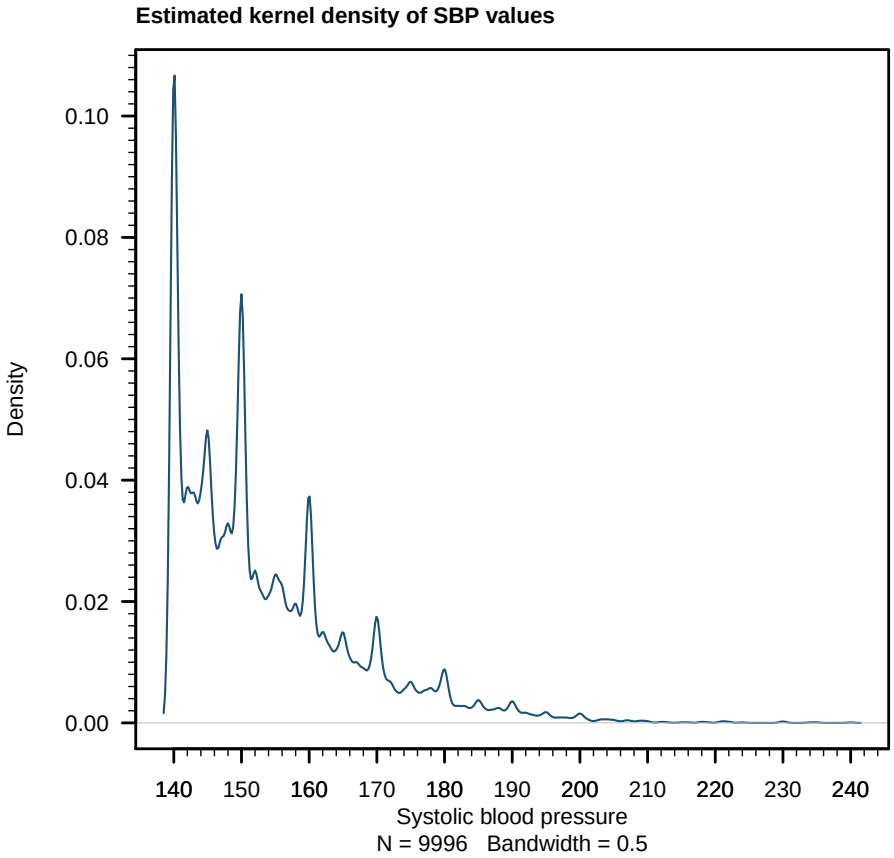

Fig. 7.5: Estimated kernel density function of the observed systolic blood pressure values for individuals who suffer from isolated systolic hypertension

district of Braga. Note that for $u = 160$ mmHg, we reject the possibility of a GPD model fit for the data, though as soon as we increase to 161 mmHg, both test statistics radically drop almost twofold and, therefore we now do not reject the GPD model. A similar phenomena happens at 169 mmHg and 170 mmHg, where the Anderson-Darling test statistic drops almost threefold.

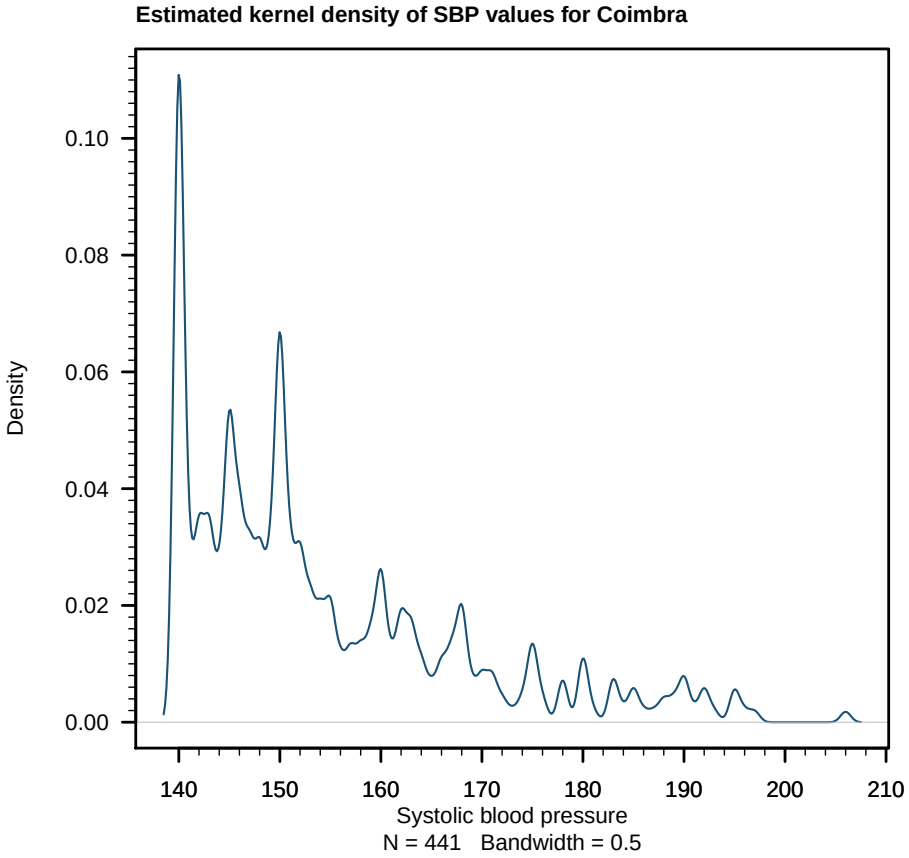

Fig. 7.6: Estimated kernel density function of the observed systolic blood pressure values for individuals who suffer from isolated systolic hypertension for the district of Coimbra

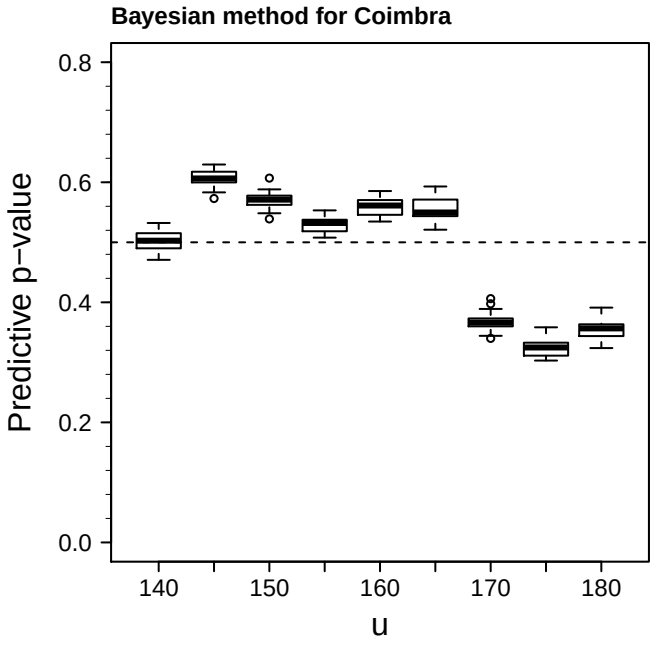

Fig. 7.7: Bayesian method for threshold selection using measures of surprise for Coimbra's observations of systolic blood pressure.

Table 7.5: Extreme quantiles for each Portuguese district and islands. Empirical estimates for $q_{0.995}$ are not included since for some districts there are few observations above this value. (*) values greater than 300 mmHg

District	Empirical$_{q_{0.99}}$	Model$_{q_{0.99}}$	IC 95%$_{q_{0.99}}$	Model$_{q_{0.995}}$	IC 95%$_{q_{0.995}}$	Max	Endpoint
Açores	198.91	198.21	(187.67, 232.37)	204.97	(192.37, 252.97)	205	287.60
Aveiro	191.65	192.68	(190.03, 196.52)	196.49	(193.49, 201.57)	204	210.96
Beja	177.52	181.28	(178.33, 189.11)	183.32	(180.39, 193.72)	186	190.20
Braga	199.00	199.11	(194.62, 206.36)	206.11	(200.39, 217.39)	221	*
Bragança	195.75	195.62	(190.23, 205.11)	200.79	(194.44, 214.96)	208	239.97
Castelo Branco	200.00	200.80	(193.16, 221.67)	207.32	(198.59, 242.77)	216	*
Coimbra	195.00	195.59	(192.75, 201.34)	199.26	(195.72, 206.73)	206	213.69
Évora	192.10	193.71	(184.59, 219.00)	201.78	(190.21, 246.19)	213	*
Faro	190.74	191.31	(186.93, 199.99)	194.98	(190.12, 206.02)	198	212.94
Guarda	186.00	186.18	(182.19, 197.63)	188.28	(184.24, 202.54)	187	194.52
Ilha da Madeira	195.60	195.09	(188.39, 210.17)	200.95	(193.22, 222.48)	206	244.15
Leiria	203.34	203.55	(197.85, 212.70)	210.52	(203.84, 223.15)	224	277.80
Lisboa	195.53	196.93	(193.90, 200.63)	204.63	(200.44, 210.31)	235	*
Portalegre	191.27	190.87	(185.16, 206.27)	194.16	(187.88, 213.89)	194	208.38
Porto	196.11	196.57	(194.18, 200.93)	203.67	(199.41, 210.28)	240	*
Santarém	204.51	205.74	(203.89, 211.53)	210.93	(209.04, 218.76)	221	226.84
Setúbal	190.39	191.90	(188.87, 196.16)	196.45	(193.00, 202.15)	210	222.21
Viana do Castelo	186.62	185.88	(179.96, 200.69)	189.91	(182.86, 209.60)	190	217.10
Vila Real	195.00	197.79	(190.97, 211.06)	204.88	(196.67, 223.86)	222	*
Viseu	203.00	205.43	(198.63, 218.43)	212.44	(204.60, 231.33)	230	276.72

Table 7.6: Extreme quantiles for each Portuguese district and islands using the exponential model. Empirical estimates for $q_{0.995}$ are not included since for some districts there are few observations above this value

District	Empirical$_{q_{0.99}}$	Model$_{q_{0.99}}$	IC 95%$_{q_{0.99}}$	Model$_{q_{0.995}}$	IC 95%$_{q_{0.995}}$
Açores	198.91	206.46	(191.21,221.70)	216.88	(199.24,234.52)
Braga	199.00	200.70	(194.82,206.57)	209.41	(202.14,216.68)
Castelo Branco	200.00	202.15	(191.56,212.75)	210.05	(197.01,223.10)
Évora	192.10	194.02	(181.82,206.21)	202.50	(187.27,217.72)
Lisboa	195.53	196.85	(193.61,200.10)	204.65	(199.81, 209.49)
Porto	196.11	196.84	(193.01,200.67)	204.42	(199.54,209.29)
Vila Real	195.00	201.60	(193.40,209.81)	210.80	(201.16,220.43)
Viseu	203.00	210.11	(199.58,220.64)	220.63	(207.88,233.37)

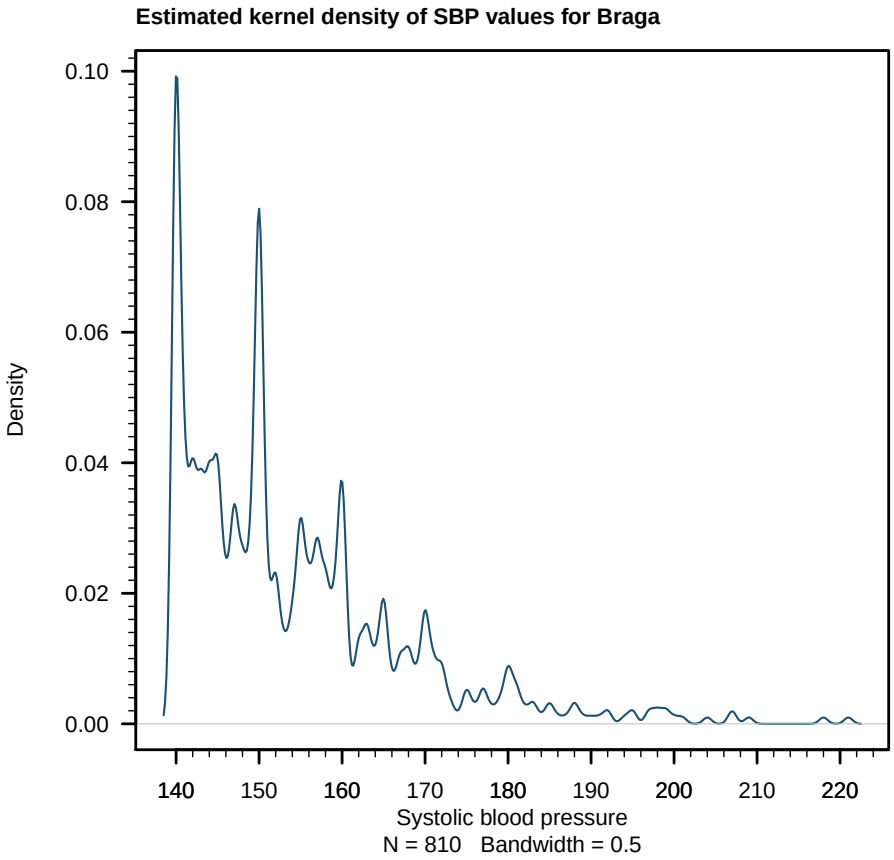

Fig. 7.8: Estimated kernel density function of the observed systolic blood pressure values for individuals who suffer from isolated systolic hypertension for the district of Braga

Tabela 7.7: Cramér-von Mises e Anderson-Darling goodness-of-fit test for the GPD for the values of SBP in Braga

u	$\hat{k}$	W^2	A^2	p_{A^2}	p_{W^2}
140	-0.1747	0.7144	5.0807	~ 0	~ 0
150	-0.1609	0.6048	4.2326	~ 0	~ 0
160	-0.1681	0.2536	2.0084	~ 0.001	~ 0
161	-0.1258	0.1422	1.1276	~ 0.025	~ 0.05
164	-0.0956	0.1222	0.9394	> 0.05	> 0.05
169	-0.1046	0.1544	1.2684	~ 0.025	~ 0.05
170	-0.2162	0.0608	0.4814	~ 0.25	~ 0.25
180	-0.2921	0.0318	0.2901	~ 0.5	~ 0.5

8 | Modeling Extreme Systolic Blood Pressure Values in the Elderly

In this chapter we propose more robust models of extremes for the values of blood pressure in elderly individuals (>=55) who suffer from isolated systolic hypertension. There are several reasons behind our interest in this study. Firstly, the exploratory analysis hinted that systolic blood pressure values somewhat change between age strata (as seen in figure 6.3), suggesting that as a person ages, his or her systolic blood pressure rises. Secondly, the elderly make up the bulk of the observations. We recall the proportions of each age stratum in table 6.3 where approximately 86% of the sample's observations were people aged 55 and higher. Out of the 9996 individuals who suffer from isolated systolic hypertension, 8174 belong to this group (there were 477 individuals with missing age value). Finally, we apply new statistical methods to deal with the issue of high absolute frequencies for certain values, the quantization structure of the data and account for the multiple testing problem.

We begin by addressing the quantized structure of the data. As mentioned in previous sections, the methods to model extreme values were constructed for continuous variables, hence some methods might not perform well when applied to a highly discretized dataset. We quote [Bader et al., 2018] regarding the performance of the goodness-of-fit tests using a quantized dataset: "Quantization pushes the null distribution of the Anderson-Darling statistic to the right; the p-value obtained by positioning the observed statistic with the quantized data to the null distribution from continuous data is smaller than it should be". Thus, we may be led to reject a certain model that in fact was fit for the data.

By making an exploratory analysis of individuals in this study (individuals aged 55 and higher and suffering from isolated systolic hypertension), we face similar issues to those mentioned during the analysis of the Portuguese districts, specifically the quantized structure of the data and the high frequencies of *rounded* numbers. Figure 8.1 illustrates this issue. The systolic blood pressure values of 140, 150, 160, 170, 180, 190 and 200 display higher frequencies than those of their neighbors. The reason for this behavior is unknown, though one might assume that it was the result of biased approximation or perhaps machinery that did not account for unit rounding. The most common method to deal with this problem is to *shake* the sample distribution, considering each value censured in an interval. For some observed value x_{obs}, its true value x belongs to an interval $[x_{obs}-\delta, x_{obs}+\delta]$, $\delta \in \mathbb{R}^+$. We can choose how x is distributed in this interval, for example, x can be equally distributed amongst this interval, or given a higher probability to be close to the observed value x_{obs}. The former can be constructed by generating a set of random values from a continuous Uniform distribution with parameters $a = -\delta$ and $b = \delta$, $\delta \in \mathbb{R}^+$, and adding them to each observed value. The latter can be obtained by generating values from a beta distribution with parameters $\alpha = \beta = \delta$, $\delta > 0$, and location parameter 0.5, hence taking values in $[-0.5, 0.5]$. This second alternative will result in a milder shakeup of the data when compared to the first, since it is more likely that the generated values will be close to 0 and have lower bound $x_{obs} - 0.5$ and upper bound $x_{obs} + 0.5$. We would like to point out that this technique is used in several studies, see [Bader et al., 2018], and is usually applied in order to obtain a smoother empirical distribution, though it is important to underline that, technically, the data is being altered, hence usually a mild jitter is considered.

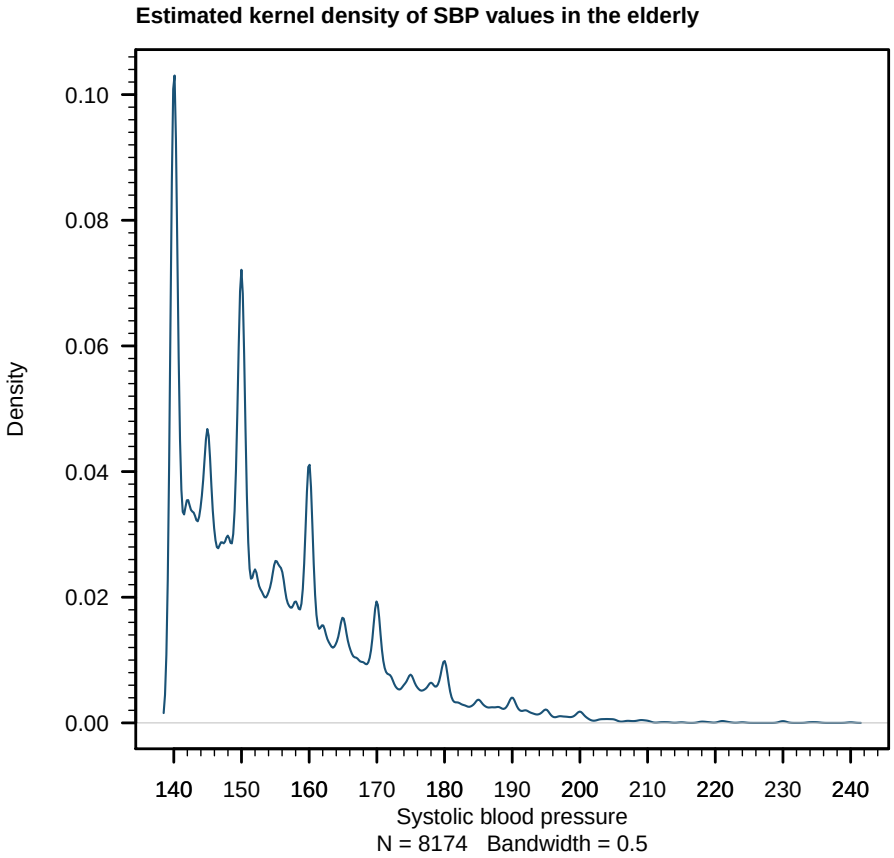

Fig. 8.1: Estimated kernel density function of the observed systolic blood pressure values for elderly individuals who suffer from isolated systolic hypertension

8.1 Jitter and Non-jitter Extreme Value Models for Systolic Blood Pressure in Individuals Who Suffer From Isolated Systolic Hypertension

We aim to produce three extreme models for systolic blood pressure for the individuals with the aforementioned criteria, using three distinct datasets.

1. Data + Beta$(10, 10, -0.5, 0.5)$,
2. Data + Uniform$(-1.5, 1.5)$,
3. Unaltered Data

We want to answer several questions by comparing the three models. Firstly, the behavior of the different threshold selection methods - does the jittering process radically change the outcome of these methods? Secondly, how does a mild jitter (first dataset) compare to a more agitated jitter (second dataset)? Thirdly, how do the resulting models hold up regarding their predicting capabilities?

Using the R function *rbeta*, we generated a random sample with size 8174 from a beta distribution with parameters $\alpha = 10$, $\beta = 10$ and location parameter 0.5.

Similar to the previous case, we generated a sample of size 8174 from the continuous uniform distribution with parameters $a = -1.5$ and $b = 1.5$ using the R function $runif$. Figures 8.2 and 8.3 represent the estimation of the kernel density functions for the simulated samples.

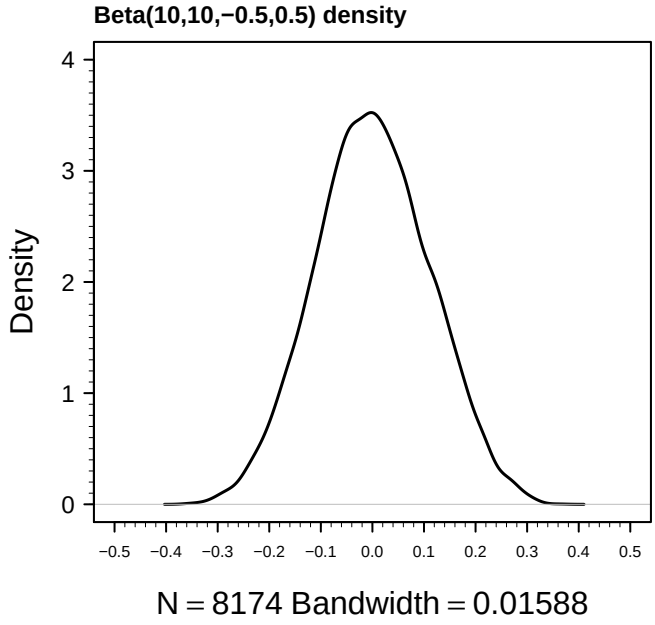

Fig. 8.2: Kernel density function estimation for a sample generated from a beta(10,10,-0.5,0.5) distribution ($n = 8174$)

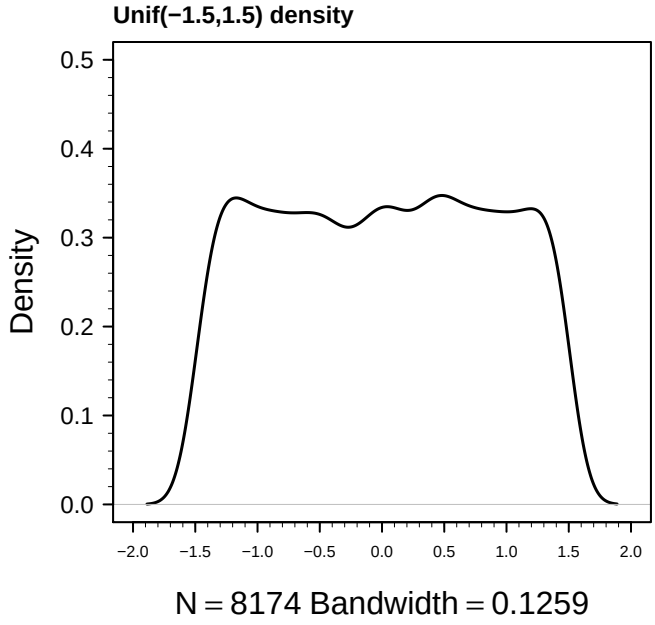

Fig. 8.3: Kernel density function estimation for a sample generated from a uniform(-1.5,1.5) distribution ($n = 8174$)

We then create two new datasets by adding each sample to the data. Note that by adding these simulated samples to the SBP values of elderly individuals who suffer from ISH, we might get some values below 140 mmHg. Those values are not considered in the subsequent analysis.

Let's now investigate how both jitters altered the data.

Figure 8.4 illustrates the histograms and density functions of the non-jitter data. Note that both jitters appear to *smooth* out the sample's distribution. This can be noticed more clearly when comparing the density function of the non-jitter data with the uniform-jitter data. Plus, the histogram shows that there is

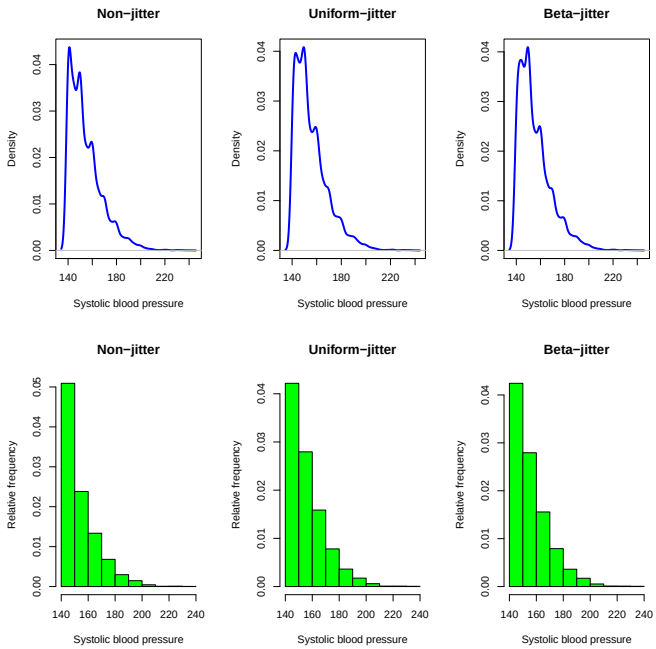

Fig. 8.4: Histograms and kernel density estimation for the non-jitter data and jitter data using the uniform and beta distributions

less frequency differences between neighboring classes. It is important to underline that although there appears to be a slight difference between the jitter data and the non-jitter data, the summary statistics of these data sets seem to not differ significantly, as seen in table 8.1.

Table 8.1: Summary of the systolic blood pressure by age in the uniform jitter-data, beta-jitter data and non-jitter data

Data	**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
Non-jitter	140.0	145.1	151.9	155.6	162.0	240.00
Unif-jitter	140.0	145.6	151.7	155.7	162.0	238.50
Beta-jitter	140.0	145.2	151.9	155.8	162.0	239.98

Next,we propose a sequence of possible threshold candidates (140,150,160,170,180,190,200). Figures 8.5, 8.6 and 8.7 present the exponential QQ-plots and histograms for the data above each candidate threshold for the non-jitter, uniform-jitter and beta-jitter data. For values above 170 mmHg the exponential model seems to adequately fit the 3 cases. Furthermore, the histograms above this value display a tail decay indicating as well that an exponential model could give an adequate fit.

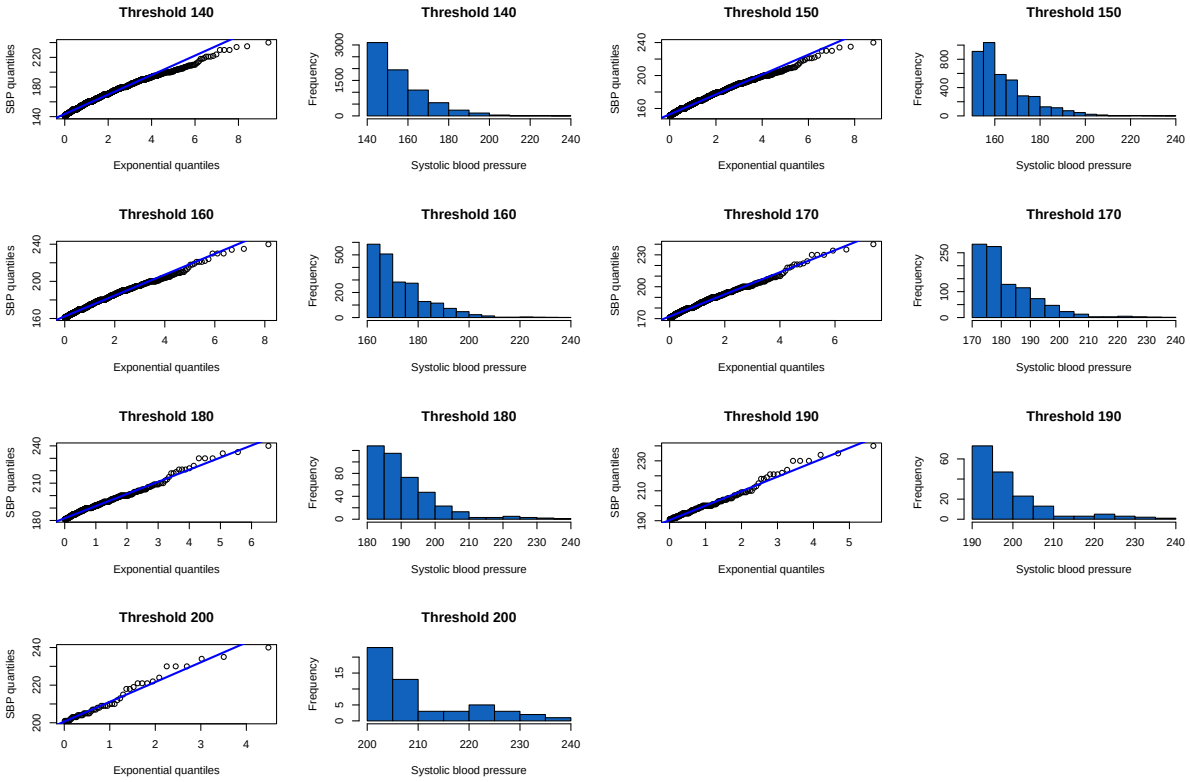

Fig. 8.5: Exponential QQ-plots and histograms for each candidate threshold for the non-jitter data

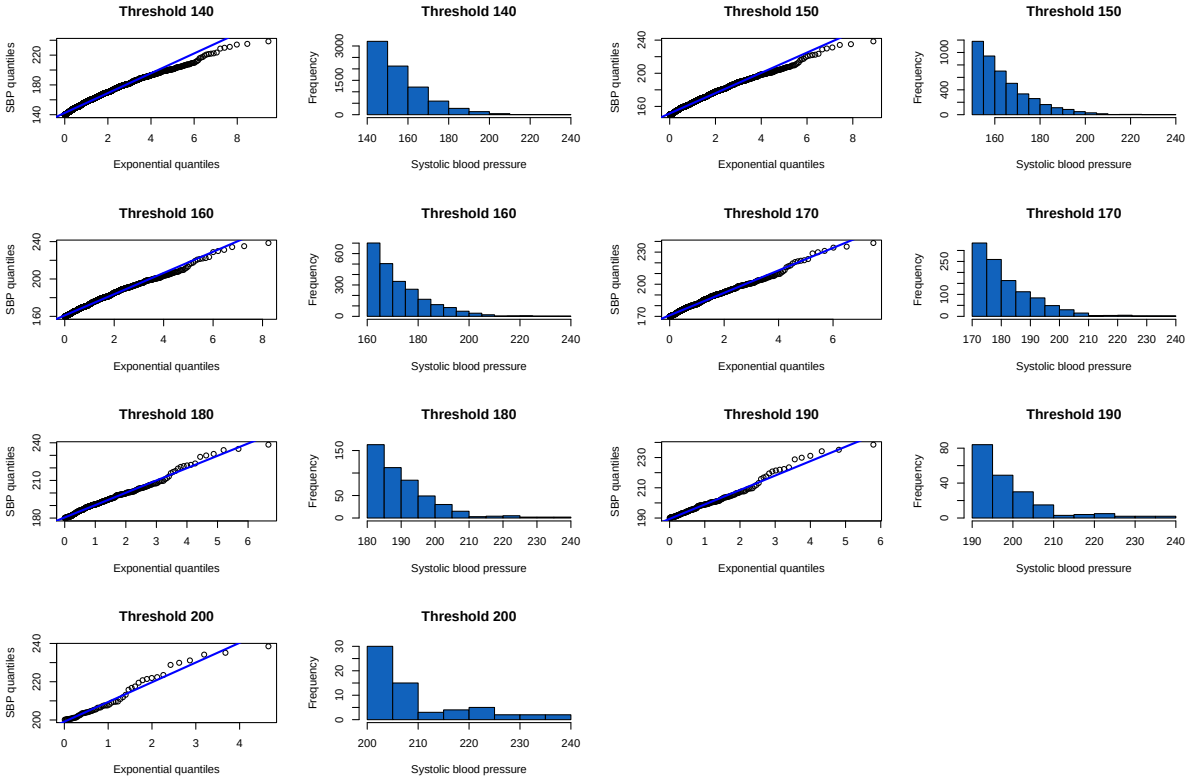

Fig. 8.6: Exponential QQ-plots and histograms for each candidate threshold for the uniform-jitter data

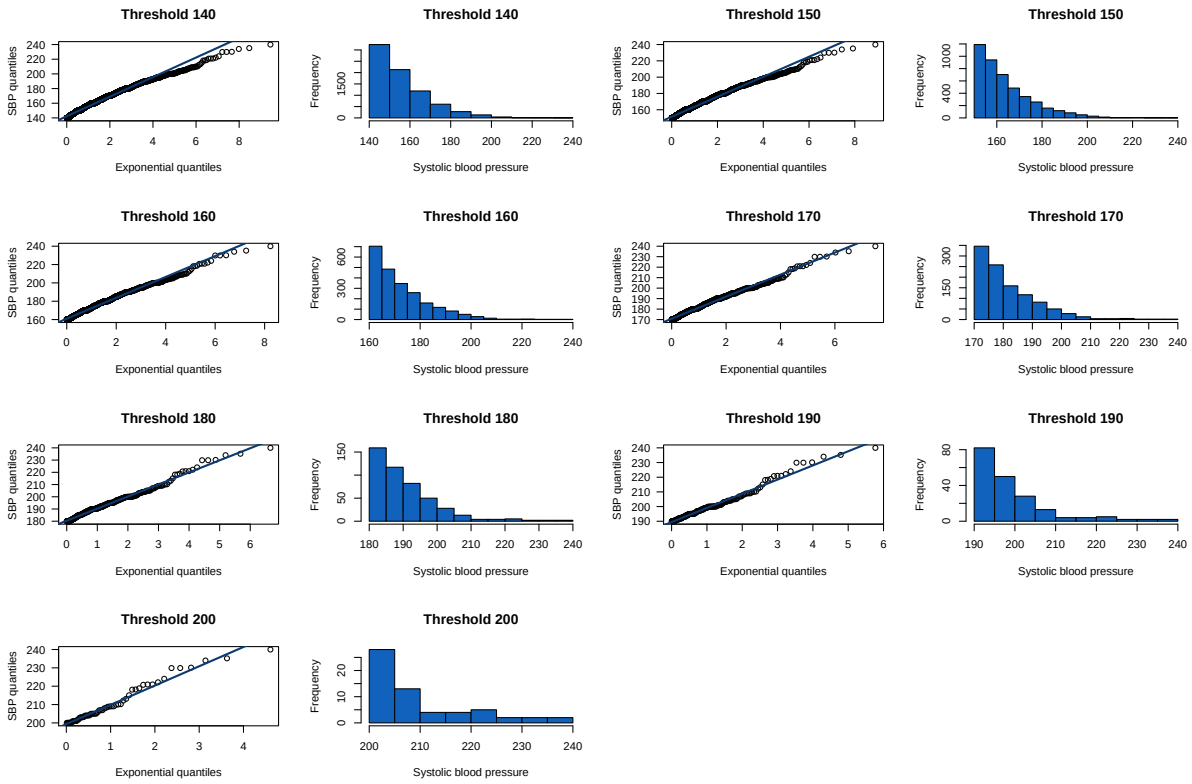

Fig. 8.7: Exponential QQ-plots and histograms for each candidate threshold for the beta-jitter data

8.2 Threshold Selection Analysis

We now start the procedure of selecting adequate thresholds for each case, with the goal of fitting GPD models for each threshold's excesses. We start with the mean residual life function. As stated before, this function should have a linear behavior for some high value of systolic blood pressure. Using the R package *exTremes* and its function *mrlplot*, we plotted this function for each data set. Figures 8.8, 8.9 and 8.10 illustrate the results.

The plot seems to indicate that for values between 180 mmHg and 200 mmHg, the function appears to have a linear-like behavior, implying that an appropriate threshold could lie between these two values.

Figure 8.11 plots the predictive p-values obtained by considering the vector of threshold candidates (140,150,160,170,180,190,200) for the non-jitter data, uniform-jitter data and beta-jitter data, respectively. For each threshold, we sampled the predictive posterior distribution 5000 times. We then proceeded to compute the p-values. This process was repeated 30 times and presented in the figure by a boxplot at each threshold. We remind the reader that each p-value obtained per threshold can be interpreted as evidence against the GPD model when it is close to 0 or 1. Furthermore, the resulting p-values can be understood as showing less incompatibility with the GPD model when near 0.5 [Meng, 1994], [Lee et al., 2015]. Figure 8.11, left, corresponds to the method applied to the non-jitter dataset. It only manifests less incompatibility with the GPD model for high threshold values, i.e., $190 < u < 200$ mmHg. Moreover, the p-values demonstrate a switch in surprise when more data is considered, i.e., when we introduce data below 190 mmHg, p-values move away from 0.5 and tend to 0, suggesting more incompatibility with the model. On the other hand, the p-values obtained for both jitter cases do not seem to change a great deal until we consider the data below 150 mmHg. This method seems sensible to the jitter process, even in the case of the mild beta-jitter, since it produces overall higher p-values in both jitter instances. Based on this output, we are led to select a high threshold value for the non jitter case, i.e., a value between 190 mmHg and 200 mmHg. Both jitter cases seem to indicate that 150 mmHg is an acceptable threshold, since there was a change in surprise from 140 mmHg to 150 mmHg, meaning

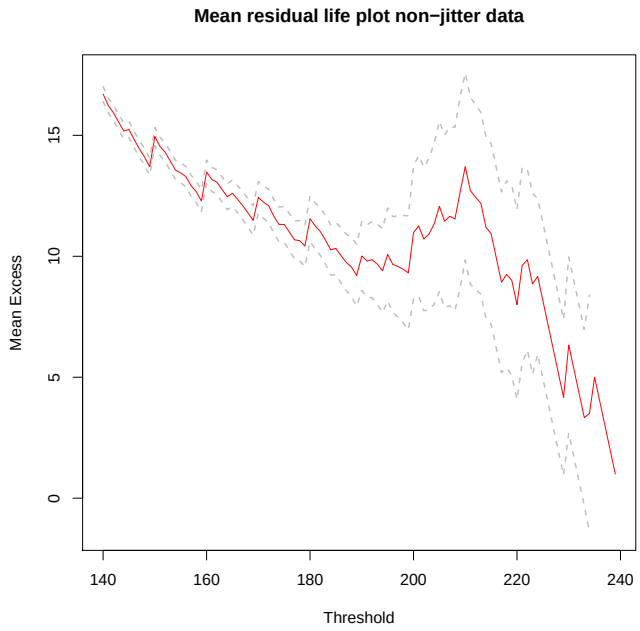

Fig. 8.8: Mean residual life function for the non-jitter data

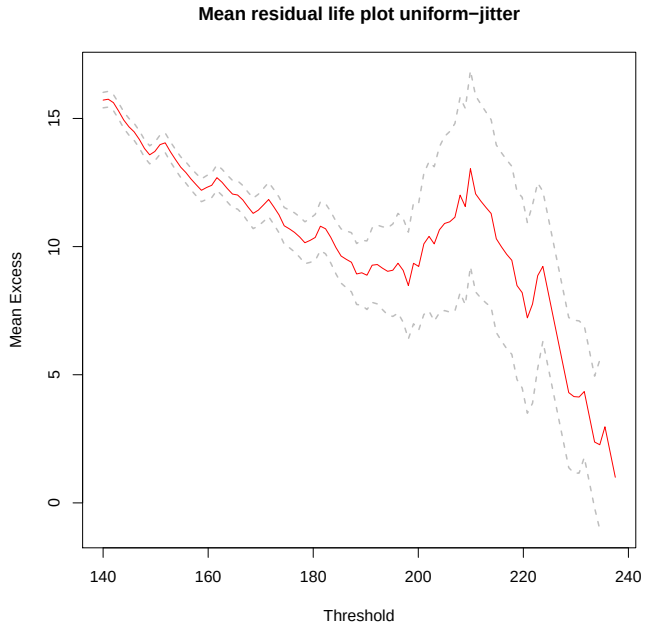

Fig. 8.9: Mean residual life function for the uniform-jitter data

the predictive p-value obtained for 150 mmHg is closer to 0.5 than the one obtained from 140 mmHg. Furthermore, for the remaining threshold values, the predictive p-values obtained do not appear to shift a great deal.

Next, we present the automated threshold selection using goodness-of-fit tests for each of the previously mentioned data sets. We will adopt the ForwardStop rule outlined in [Bader et al., 2018] and [G'Sell et al., 2015]. Let $u_1, u_2, ..., u_m$ be a sequence of candidate thresholds for a given data set, and the order test hypotheses $H_0^1, H_0^2, ..., H_0^m$, where for some $1 \leq i \leq m$, H_0^i : the excesses over u_i come from a generalized Pareto distribution. Let $p_1, p_2, ..., p_m$ be the p-values obtained using the Cramér-von Mises

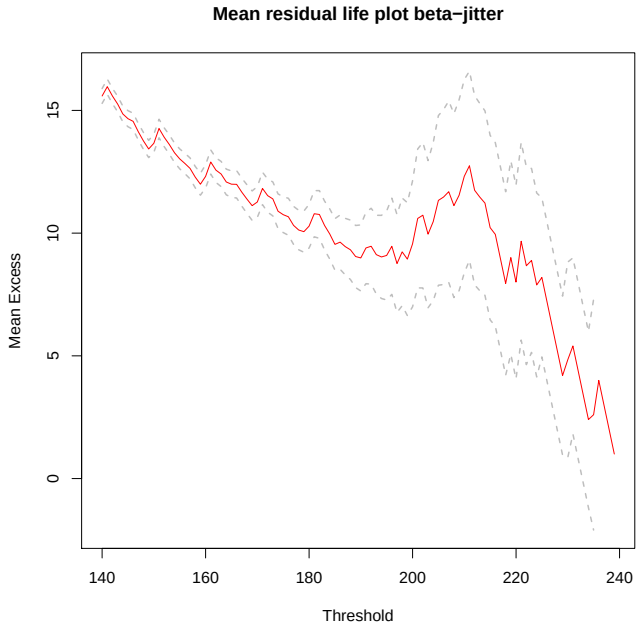

Fig. 8.10: Mean residual life function for the beta-jitter data

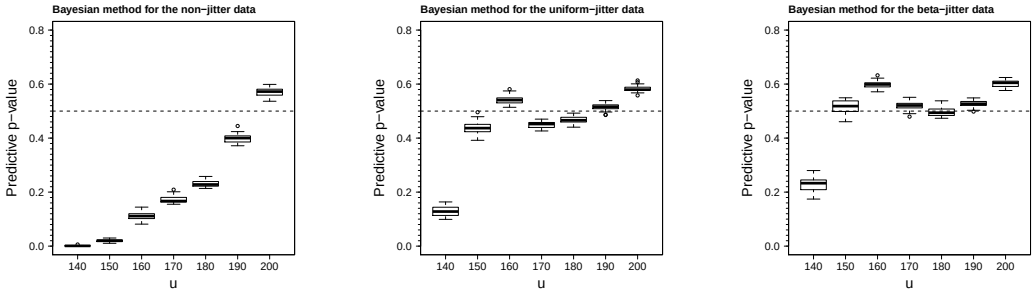

Fig. 8.11: Bayesian threshold selection method using measure of surprise for the non-jitter, uniform-jitter and beta-jitter data

goodness-of-fit test for each sample. The FowardStop rule is given by:

$$\hat{i} = \max\left\{ i \in \{1,...,m\} : -\frac{1}{i}\sum_{j=1}^{i} log(1-p_j) \leq \alpha \right\}, \tag{8.1}$$

where α is the prior set significance level. This rule has been shown to control the error rate associated with wrongly accepting a low threshold. The simple outline of this method is to compute the p-values at each threshold, starting from the smallest and computing this mean until (8.1) can be satisfied. Once $\hat{i}$ is obtained, we reject H_i for $i = 1,...,\hat{i}$, thereby not rejecting the null hypothesis at $\hat{i}+1$ and accepting the threshold associated with $H_0^{\hat{i}+1}$ as the adequate selection.

Table 8.2: Results of the automated threshold selection using the Cramér-von Mises goodness-of-fit tests for the non-jitter dataset

threshold	**num.above**	***p*-values**	**fowardstop**	**statistic**
140	7113	2.4221e-47	$\sim$0	3.3111
150	4012	1.3233e-46	$\sim$0	5.3901
160	2065	1.1008e-06	3.66926e-07	0.6684
170	973	1.4698e-05	3.9497e-06	0.5456
180	416	1.4609e-03	2.9556e-04	0.3260
190	173	6.8810e-02	1.2128e-02	0.1320
200	53	1.6762e-01	3.6604e-02	0.0988

Table 8.2 illustrates the results of the FowardStop rule for the non-jitter data using the R package *eva* as outlined in [Bader et al., 2018]. The results show that we should reject the first five hypotheses, at $\alpha = 0.01$ and select 190 mmHg as the adequate threshold, since the fifth test is the last test where the FowardStop mean is still below 0.01. We would like to point out that these results are in accordance with the results obtained from the Bayesian threshold selection method. Table 8.3 shows the results of the FowardStop rule for the uniform-jitter dataset. Here, the rule proposes a lower threshold. 190 mmHg is the first threshold that produces a *p*-value above 0.01. However, this *p*-value is much larger than 0.01, which might suggest that a proper threshold might lie between 180 mmHg and 190 mmHg.

Table 8.3: Results of the automated threshold selection using the Cramér-von Mises goodness-of-fit tests for the uniform jitter dataset

threshold	**num.above**	***p*-values**	**fowardstop**	**statistic**
140	7593	1.4383e-67	$\sim$0	7.9165
150	4391	2.5836e-08	1.2918e-08	0.8708
160	2269	1.4901e-03	4.9709e-04	0.3264
170	1064	2.0113e-03	8.7615e-04	0.3118
180	471	3.5537e-03	1.4129e-03	0.2763
190	196	5.7393e-01	1.4337e-01	0.0496
200	63	1.5124e-01	1.4631e-01	0.0960

Table 8.4: Results of the automated threshold selection using the Cramér-von Mises goodness-of-fit tests for the beta-jitter dataset

threshold	**num.above**	***p*-values**	**fowardstop**	**statistic**
140	7628	5.2077e-25	$\sim$0	2.8071
150	4391	1.9948e-18	$\sim$0	1.9746
160	2260	5.7836e-09	1.9279e-09	0.9210
170	1072	1.2669e-06	3.1817e-07	0.6539
180	468	1.3996e-04	2.8249e-05	0.4265
190	227	1.2325e-01	2.1946e-02	0.1095
200	74	1.8077e-01	4.7295e-02	0.0930

Table 8.4 shows the FowardStop rule results for the beta-jitter dataset. Here, we reject the first 5 hypotheses at α=0.01, thus suggesting that 190 mmHg is an adequate threshold value. The test statistics from the beta-jitter aren't as inflated as the original dataset.

Only the Cramér-von Mises test was applied, since the test statistic results obtained from the Anderson-Darling test for these data sets were highly inflated.

We've selected $u = 190$ mmHg as the threshold for the three cases. Next, we present the fitted models for each dataset.

Table 8.5: Extreme models for the non-jitter, beta-jitter and uniform-jitter data. $(*)$ the support does not have an upper finite boundary

Model	u	n	$\hat{k}$	95% CI for k	σ	95% CI for σ	max	endpoint	$-\ln(L)$
Non-jitter	190	173	-0.049	(-0.190;0.093)	10.50	(8.34;12.65)	240	406.08	571.34
Unif-jitter	190	196	0.062	(-0.097;0.222)	8.37	(6.60;10.15)	238.50	$*$	624.78
Beta-jitter	190	192	0.062	(-0.100;0.224)	8.47	(6.65;10.29)	239.98	$*$	614.13

The results presented in table 8.5 were obtained using the *gpd.fit* function from the *ismev* R package. It seems as though the non-jitter dataset for exceedances over 190 mmHg produced a light tail model, where the estimated shape parameter is negative and although its maximum likelihood 95% confidence interval contains 0, it is heavily skewed to the negative side of the axis. The other two offered a more exponential-like tail model, since the estimates obtained for k in each jitter dataset are close to 0, though their intervals are slightly skewed to the positive side of the axis.

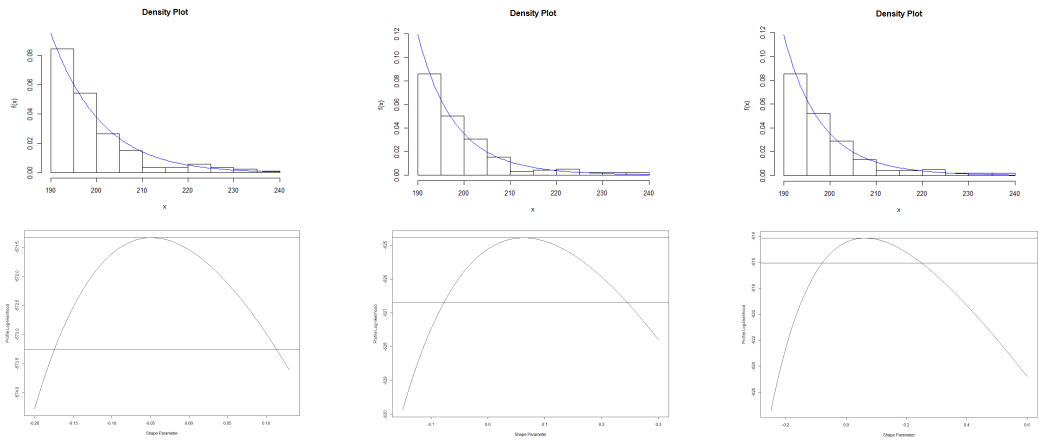

Fig. 8.12: Density plots with histogram and profile likelihood plots for the non-jitter (left), uniform-jitter (center) and beta-jitter (right) function

Figure 8.12 presents the histograms with density function for each model and the profile likelihood plots with 95% confidence intervals. Both jitter models appear to adequately fit the data. The non-jitter model only fits the data fairly, displaying the apparent difficulties of fitting a continuous model to such a highly discretized data set. The profile likelihood plots yield confidence intervals similar to those obtained via the asymptotic properties of the maximum likelihood estimators.

For each of the considered models, we will test if there is a statistical significance in considering the more parsimonious exponential model, which has scale parameter σ, instead of the GPD model, which has the parameter vector (k, σ), where k is the shape parameter. For each model, our hypotheses consist in $H_0 : k = 0$ *vs.* $H_1 : k \neq 0$. We use the deviance function as the test statistic:

$$T = 2\big(l_{M_1}(\mathbf{x}) - l_{M_2}(\mathbf{x})\big) \sim \chi^2_1,$$

where, in this case, $l_{M_1}(\mathbf{x})$ is the log-likelihood function for the GPD model and $l_{M_2}(\mathbf{x})$ is the log-likelihood function for the exponential model. We present the results in table 8.6.

Table 8.6: Results of the deviance test for non-jitter model, uniform-jitter model and beta jitter-model

Model	l_{M_1}	$l_{M_2}(x)$	T	p
Non-jitter	-571.3383	-571.7379	0.7992268	0.3713246
Unif-jitter	-624.7814	-625.5246	1.486305	0.2227907
Beta-jitter	-614.1304	-614.8442	1.427657	0.2321473

At the usual significance levels, we do not reject the null hypothesis for the three models. There is no evidence that the exponential model is not suited for the data.

Using the formulas (3.17) and (3.18) obtained in a previous chapter, we can calculate some tail probabilities and extreme quantiles. Extreme quantiles in the extreme value analysis setting can be thought of as return levels. The systolic blood pressure return level that is exceeded in mean once every ten thousand people corresponds to the model quantile $\chi_{0.9999}$. Using the previous formulas, we calculate this value for each considered model.

The exponential model was selected as the most parsimonious in the three cases.

For the three datasets, the exponential model was selected. By using the deviance test, we proved that there wasn't a significant difference between the exponential model and the GPD model. We consider the following formula obtained in a previous chapter to calculate extreme quantiles for the exponential model.

$$x_p = \sigma \ln\left(\frac{\tau_u}{p}\right) + u. \tag{8.2}$$

Note that p needs to be low enough such that $x_p - u > 0$. We compute the value that is exceeded in mean once every ten thousand individuals for each model. The values obtained for each model were as follows: for the uniform-jitter model, we obtained $x_{\frac{1}{10000}} = 239.5882$ mmHg, the beta-jitter model provided $x_{\frac{1}{10000}} = 239.9046$ mmHg and, finally, for the non-jitter model, we obtained $x_{\frac{1}{10000}} = 243.6111$ mmHg. The results obtained for each jitter case are similar, but differ substantially from the value obtained for the non-jitter model.

Table 8.7 presents some more extreme quantiles, this time using the GPD models for the three jitter cases. These were obtained using the *gpd* and *tailplot* functions from the R package *evir*. Table 8.8 presents the same extreme quantiles, this time using the exponential model, since it was deemed more parsimonious than the GPD model in the three cases. In both tables, we compare some empirical quantiles with model quantiles. We note that empirical quantile estimation was obtained using the R function *quantile* and for any given quantile there were at least 8 observations above the estimated value. The uniform-jitter model and the beta-jitter model supplied highly accurate quantile predictions when compared to the empirical distribution. Moreover, in both cases, the 95% confidence interval provided by the model for the 0.99 quantile contained the empirical estimate of set quantile. In medical terms there is no distinction between the results obtained for the three methods.

Table 8.7: Extreme quantiles for the uniform-jitter model, beta-jitter model and non-jitter model

Model	Empirical$_{q_{0.99}}$	Model$_{q_{0.99}}$	IC 95%$_{q_{0.99}}$	Empirical$_{q_{0.995}}$	Model$_{q_{0.995}}$	IC 95%$_{q_{0.995}}$
Non-jitter	198	197.73	(196.09,199.37)	203	204.62	(201.28,207.97)
Unif-jitter	198.63	198.18	(196.34,200.03)	203.98	204.48	(200.85,208.10)
Beta-jitter	198.69	198.05	(196.21,199.88)	203.95	204.40	(200.76,208.04)

We expected the non-jitter fit not to deliver such great results. It delivers an excellent fit to the data. One could argue that the high absolute frequency problem dissipates for the high systolic blood pressure values considered in the data used to create this model, thus resulting in an adequate fit.

Table 8.8: Extreme quantiles for the uniform-jitter model, beta-jitter model and non-jitter model using the exponential model

Model	Empirical$_{q_{0.99}}$	Model$_{q_{0.99}}$	IC 95%$_{q_{0.99}}$	Empirical$_{q_{0.995}}$	Model$_{q_{0.995}}$	IC 95%$_{q_{0.995}}$
Non-jitter	198.00	197.51	(196.38,198.63)	203	204.45	(202.29,206.60)
Unif-jitter	198.63	198.47	(197.28,199.66)	203.98	204.66	(202.60,206.71)
Beta-jitter	198.69	198.33	(197.15,199.52)	203.95	204.59	(202.52,206.66)

As mentioned before, the models appear not to differ a great deal. Future work could be developed by considering stronger *jittering* methods which may result in a different threshold selection and, thus, provide contrasting models.

9 | Comments, Conclusions and Future Work

This dissertation's main objective was to apply the *Peaks Over Threshold* methodology to create models for several instances of systolic blood pressure in individuals who suffer from isolated systolic hypertension, with the purpose of extrapolating on the levels of SBP in the most severe cases of this disease. The data consisted in a sample of 40065 voluntary individuals who attended a Portuguese pharmacy during the campaign organized by the National Portuguese Pharmacy Association. Several biometric variables were recorded for each individual, such as total cholesterol, systolic blood pressure, diastolic blood pressure, body mass index and others. No information was included about an individual's medication history. The individuals who attended the campaign did so voluntarily, thus, this sample does not represent the Portuguese population - it only reflects the population that attends these pharmacies.

In chapters 2 and 3 we discuss the layout of Extreme Value Theory. Here, we derive the framework for the models obtained in later chapters. We felt that writing chapter 2 was key to better understand the intricacies of EVT. In chapter 3, we also address several threshold selection methods, including a state-of-the-art Bayesian procedure, goodness-of-fit tests and other classical threshold selection techniques. Moreover, the interest was also to find out if all the threshold selection techniques were able to capably point towards an equal threshold candidate. Section 3.5 uses the delta method presented in chapter 5 to obtain confidence intervals for extreme quantiles.

Chapter 4 presents a brief summary of Bayesian statistics. The goal here is to present an beginner's guide so that the reader can better follow the Bayesian method presented in chapter 3.

Chapter 5 addresses, with some detail, a mathematical procedure which enables the calculation of confidence intervals for extreme quantiles. We felt it was necessary to include this chapter in order to help the reader better understand techniques that use this procedure.

Chapter 6 introduces the reader to the exploratory analysis of the data. The objective of this inquiry is to profile individuals in terms of SBP values who suffer from ISH, according to several measures of interest, such as age, tobacco consumption, BMI, gender and district. Studying the behavior of SBP by age stratum showed compelling results. It suggested that as an individual aged, his or her SBP rose. This relation is well known in the literature, see [Pinto, 2007] and [Bavishi et al., 2016]. This result suggests that a model accounting for this variable could be created, such as a regression model. This result was also key in creating the model described in chapter 8. Gender, tobacco consumption and BMI displayed little to no effect in SBP values between each stratum. Regarding the geographical variables, higher extremes of SBP were observed in more densely populated districts.

From this chapter onwards we consider a subsection of individuals as our group of interest - those who suffer from isolated systolic hypertension (SBP value equal or above 140 mmHg, and DBP lower than 90 mmHg). There were several reasons behind our interest in this group. Firstly, this group constitutes the bulk of the hypertensive individuals (9996 individuals). Secondly, in chapter 8, we fit a model of extremes to elderly individuals (more than 55 years old), which have been shown to have high incidence in this category of hypertension, see [Bavishi et al., 2016].

Before deciding to explore the high values of SBP in individuals who suffer from ISH, we considered two other possible hypertension cases. First, individuals who have SBP$<$ 140 mmHg and DBP$>$ 90.

This group constituted a small partition of the sample (1076 individuals), hence we decided that it was not worthwhile to explore modeling of high values of DBP. The second case was constituted by the individuals who have both blood markers above than the standards, see table 6.1. Analyzing the extreme values of both SBP and DBP variables requires a more sophisticated EVT framework. Several methodologies are available in the literature to explore bivariate extreme value analysis, but are beyond the scope of this thesis, see [Heffernan and Tawn, 2004].

In chapter 7, we fit GPD models to the extremes of the SBP for each Portuguese district and island in individuals who suffer from ISH. We also discuss some of the difficulties associated with the EVT analysis performed and with the data. We use Braga as an example of a district that yielded an easy threshold selection analysis and Coimbra as an opposite case.

In this chapter we do not account for the multiple testing problem when using goodness-of-fit tests yet. It was also in this first analysis that we discovered an underlying issue with the data - its quantized structure, which can constitutes an problem, since EVT was developed for continuous variables. Even though the blood pressure is a continuous variable, the common instruments that measure the blood pressure (in mmHg) give rounded values. Most likely, more precise readings are not needed. However, this can be a real problem when using the POT methodology using discretized data. Also, *rounded* SBP values i.e., values ending in zero, such as 140, 150, 160, displayed much higher frequencies than their neighboring values. In chapter 8, we propose a solution to both of these issues by considering two different jittering models.

As mentioned above in chapter 8, we obtain models for the extreme values of SBP in elderly individuals (at least 55 years old) that suffer from ISH. Our aim here was twofold. First, elderly individuals constitute over 80% of all cases of ISH in this dataset. Second, the SBP of these individuals seems to grow as age increases, as seen in figure 6.3. The quantization and high frequency issues pointed out above are addressed by considering three different datasets. The original data, the data with an added jittering beta distribution and the data with an added stronger jittering process that uses the uniform distribution, see [Bader et al., 2018]. The multiple testing issue is addressed by considering the FowardStop rule proposed by [Bader et al., 2018] and [G'Sell et al., 2015].

Preliminary analysis of the resulting jitter datasets demonstrates that we have been successful in *breaking* the discrete feature of the recorded data, see figure 8.4. Moreover, the jittering process did not alter the data a great deal, as described in table 8.1.

As previously referred, threshold selection for each case was performed by computing the mean residual life function [Coles, 2001], the GPD goodness-of-fit Cramér-von Mises test with the FowardStop rule to account for the multiple ordered testing [Bader et al., 2018], [Choulakian and Stephens, 2001] and the Bayesian method using measures of surprise [Lee et al., 2015]. The analysis performed on the elderly individuals suffering from ISH shows that the jittering process did not alter the data to a great extent and that an appropriate threshold should lie between 180 mmHg and 190 mmHg, since it appears that in this interval the function has a linear-like behavior. The results for the Bayesian method using predictive p-values suggest threshold values for the jitter cases (u=150 mmHg) lower than the ones obtained for the non-jitter data. For the non-jitter case, an adequate threshold should lie between 190 mmHg and 200 mmHg. The Cramér-von Mises goodness-of-fit test using the FowardStop rule to account for the multiple testing problem displayed similar results for the beta-jitter and non-jitter cases, suggesting a threshold of 190 mmHg at $\alpha = 0.01$. For the uniform-jitter case, the rule suggested a value between 180 mmHg and 190 mmHg, since between these two SBP values, the test statistic dropped almost fivefold (tables 8.2, 8.3 and 8.4 display these results).

The threshold $u = 190$ mmHg was ultimately selected and, subsequently, the models were fitted to the data. Table 8.5 displays the estimated parameters for the model. Although the fitted k is negative for the non-jitter data and positive for both cases of jitter data, all the values are very close to zero, reflecting an exponential tail. In fact, the 95% confidence intervals for the shape parameter, in each case, include 0, which can also be observed in the confidence intervals obtained via the profile-likelihood function (see figure 8.12). This conjecture is further investigated by applying the deviance test. The results indicate that there are no significant differences between the GPD and the exponential distribution for each case,

as displayed in table 8.6. There is no clear difference between the predictive capabilities of each model. Future work could be developed where other jittering distributions might be used. One example could be applying a stronger jitter to the values with higher than normal absolute frequencies and a milder jitter to the remaining data.

References

[Apostol, 1967] Apostol, T. M. (1967). *Calculus: One-Variable Calculus, with an Introduction to Linear Algebra*. John Wiley & Sons, New York.

[Bader et al., 2018] Bader, B., Yan, J., and Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *Ann. Appl. Stat.*, 12(1):310–329.

[Balkema and de Haan, 1974] Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *Ann. Probab.*, 2(5):792–804.

[Bavishi et al., 2016] Bavishi, C., Goel, S., and H. Messerli, F. (2016). Isolated systolic hypertension: An update after sprint. *The American Journal of Medicine*, 129(12):1251–1258.

[Casella and Berger, 2002] Casella, G. and Berger, R. L. (2002). *Statistical inference (2nd ed.)*. Duxbury/Thomson Learning, Pacific Grove, Calif.

[Castillo and Hadi, 1997] Castillo, E. and Hadi, A. S. (1997). Fitting the generalized pareto distribution to data. *Journal of the American Statistical Association*, 92(440):1609–1620.

[Choulakian and Stephens, 2001] Choulakian, V. and Stephens, M. A. (2001). Goodness-of-fit tests for the generalized pareto distribution. *Technometrics*, 43(4):478–484.

[Coles, 2001] Coles, S. (2001). *An introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London.

[de Zea Bermudez and Kotz, 2010a] de Zea Bermudez, P. and Kotz, S. (2010a). Parameter estimation of the generalized pareto distribution—part i. *Journal of Statistical Planning and Inference*, 140(6):1353–1373.

[de Zea Bermudez and Kotz, 2010b] de Zea Bermudez, P. and Kotz, S. (2010b). Parameter estimation of the generalized pareto distribution—part ii. *Journal of Statistical Planning and Inference*, 140(6):1374 – 1388.

[de Zea Bermudez and Mendes, 2012] de Zea Bermudez, P. and Mendes, Z. (2012). Extreme value theory in medical sciences: Modeling total high cholesterol levels. *Journal of Statistical Theory and Practice*, 6(3):468–491.

[DuMouchel, 1983] DuMouchel, W. H. (1983). Estimating the stable index α in order to measure tail thickness: A critique. *Ann. Statist.*, 11(4):1019–1031.

[Grimshaw, 1993] Grimshaw, S. D. (1993). Computing maximum likelihood estimates for the generalized pareto distribution. *Technometrics*, 35(2):185–191.

[G'Sell et al., 2015] G'Sell, M. G., Wager, S., Chouldechova, A., and Tibshirani, R. (2015). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(2):423–444.

[Gumbel, 1935] Gumbel, E. (1935). Les valeurs extrêmes des distributions statistiques. *Annales de l'institut Henri Poincaré*, 5(2):115–158.

[Hajar, 2016] Hajar, R. (2016). Framingham contribution to cardiovascular disease. *Heart Views: The official Journal of the Gulf Heart Association*, 17(2):78–81.

[Heffernan and Tawn, 2004] Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

[Hosking and Wallis, 1987] Hosking, J. R. M. and Wallis, J. F. (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3):339–349.

[Kass and Raftery, 1995] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

[Lee et al., 2015] Lee, J., Fan, Y., and Sisson, S. (2015). Bayesian threshold selection on extremal models using measures of surprise. *Computational Statistics And Data Analysis*, 85:84–99.

[Meng, 1994] Meng, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.*, 22(3):1142–1160.

[Pickands, 1975] Pickands, J. I. (1975). Statistical inference using extreme order statistics. *Ann. Statist.*, 3(1):119–131.

[Pinto, 2007] Pinto, E. (2007). Blood pressure and ageing. *Postgraduate Medical Journal*, 83(976):109–114.

[Scarrott and MacDonald, 2012] Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT - Statistical Journal*, 10(1):33–60.

[Schröder et al., 2003] Schröder, H., Marrugat, J., Elosua, R., and Covas, M. I. (2003). Relationship between body mass index, serum cholesterol, leisure-time physical activity, and diet in a mediterranean southern-europe population. *British Journal of Nutrition*, 90(2):431–439.

[Turkman et al., 2018] Turkman, M. A. A., Paulino, C. D., Murteira, B., and Silva, G. L. (2018). *Estatística Bayesiana*. Fundação Calouste Gulbenkian, Lisboa.

[Wakabayashi, 2004] Wakabayashi, I. (2004). Relationships of body mass index with blood pressure and serum cholesterol concentrations at different ages. *Aging Clinical and Experimental Research*, 16(6):461–466.